Working Paper Series
Congressional Budget Office
Washington, D.C.

An Evaluation of Using Linked Survey and Administrative Data to Impute Nonfilers to the Population of Tax Return Filers

Shannon Mok
Congressional Budget Office
Shannon.Mok@cbo.gov

Working Paper 2017-06

September 2017

The author thanks Michael Berning and Joey Morales at the Census Bureau and Amy O'Hara (formerly at the Census Bureau) for their assistance in accessing the data and Edward Harris, Janet Holtzblatt, and Jim Nunns for their helpful comments. Numbers in the text and tables may not add up to totals because of rounding.

Contents

Abstract

Administrative tax return data are increasingly used for policy analysis and economic research. A potential weakness of that data source is that not everyone is required to file a tax return, even though information on the characteristics of those nonfilers is desirable for the analysis of various tax policies and tax administration. In this paper, I use data from the Census Bureau's Current Population Survey (CPS) linked to administrative tax return data to obtain demographic and income characteristics of filers and nonfilers. Those linked data are also used to model an individual's filing decision. In the absence of linked data, researchers rely on statistical matches of publicly available data—typically from the CPS and a sample of tax returns—to simulate filers and nonfilers in the population. I evaluate two statistical matches on the basis of how similar simulated filers and nonfilers are to filers and nonfilers in the linked data. The first method statistically matches records from the CPS and a public use file of tax returns by predicted income, and the second method uses the predicted probability of filing. I find that income and demographic characteristics for simulated filers under both methods are generally similar to those of filers in the linked data, but larger differences in income appear between simulated nonfilers and nonfilers in the linked data. Both simulation methods result in simulated nonfilers who have lower income than nonfilers in the linked data, although nonfilers simulated using the predicted probability method had higher income, on average, than those simulated using the predicted income method.

Keywords: microsimulation, administrative data

JEL Classification: C81, H20

Introduction

Identifying individuals who do not file federal income tax returns is challenging, yet understanding the characteristics of that population is necessary for both policy analysis and tax administration. Individuals whose income is below a certain amount (referred to as the filing threshold) generally are not required to file tax returns, although some do—typically, to pay self-employment income taxes, to obtain refunds of income taxes that were overwithheld during the year, or to receive refundable tax credits. Many other nonfilers are noncompliant individuals whose income is above the filing threshold. Understanding the characteristics of nonfilers is useful in analyzing policy options that might extend tax benefits to people outside the income tax system or that might effectively raise the filing threshold. More information on the traits of nonfilers would also be useful to the Internal Revenue Service (IRS) for the development and implementation of its compliance initiatives. In addition, research that uses administrative tax data to study issues such as income inequality and intergenerational mobility needs to account for people who do not file tax returns.[1]

Survey data—such as the Current Population Survey (CPS)—that are based on nationally representative samples of the population either do not ask respondents whether they filed income tax returns or do not allow researchers to determine on the basis of the income data collected whether an individual is required to file or has another reason for filing.[2] To overcome that limitation, some researchers—including tax modelers at the Congressional Budget Office and the Urban-Brookings Tax Policy Center—statistically match survey data to a sample of tax returns to represent the population. Records from the survey data that are statistically matched to tax records are classified as filers, and the remaining unmatched records are identified as nonfilers. Because the statistical match assigns information from multiple sources to the same individuals,

---

[1] For more details on the limitations of administrative tax return data, see Joel Slemrod, "Caveats to the Research Use of Tax-Return Administrative Data," *National Tax Journal*, vol. 69, no. 4 (December 2016), pp. 1003–1020, http://tinyurl.com/ya27wran.

[2] The Survey of Income and Program Participation asks respondents whether they filed a federal tax return in a topical module. Although response rates to qualitative questions such as "Did you file taxes?" are high, the share of survey participants responding to questions about specific tax provisions (such as whether the taxpayer itemized or claimed the earned income tax credit) are much lower. See Jeff Sisson and Kathleen Short, "Measuring and Modeling Taxes in the Survey of Income and Program Participation" (paper presented at the 2001 Joint Statistical Meetings of the American Statistical Association, Atlanta, Ga., August 5–9, 2001), https://go.usa.gov/xR7AY (69 KB).

the characteristics associated with filers and nonfilers in the resulting data set can vary from the characteristics of actual filers and nonfilers.

In recent years, the Treasury Department's Office of Tax Analysis and the staff of the Joint Committee on Taxation have developed an alternative approach that relies on a nationally representative sample of information returns filed by employers and other payers of income to construct the population of filers and nonfilers.[3] Individuals with information returns that do not have a match in the database of income tax returns are classified as nonfilers, and individuals with tax returns are filers. Because information returns are filed for an individual and contain no demographic information, analysts need to aggregate the unmatched information returns to construct tax units for married couples and individuals who can be claimed as dependents. Information returns are not publicly available, which limits the use of that method by other researchers.

Another alternative is an exact match of tax return data to survey data. The Census Bureau creates a unique data set that links survey respondents in the CPS to any federal individual income tax return (commonly known as a 1040) on which they appear as a primary or secondary filer.[4] That data set, known as the "linked data," provides researchers a way to identify filers and nonfilers, but its use is restricted to certain researchers employed at the Census Bureau or with Special Sworn Status at the agency. Individuals in the CPS who have a 1040 are identified as filers, and individuals without a 1040 in the linked data are nonfilers. Although results derived from that data—such as descriptive statistics of groups of filers and nonfilers or a model estimating the probability of filing—can be made public, the filing decision of a specific individual remains confidential.

In this paper, I evaluate two alternatives for simulating a tax unit's filing behavior to target the demographic and income characteristics of filers and nonfilers for tax year 2006, the most recent year for which publicly available tax return data were readily available at the time of this

---

[3] See Joint Committee on Taxation, *Estimating Changes in the Federal Individual Tax Model: Description of the Individual Tax Model*, JCX-75-15 (April 20, 2015), https://go.usa.gov/xR7A2 (861 KB).
[4] Those returns include Forms 1040, 1040A, and 1040EZ. In this paper, I refer to them collectively as 1040s.

analysis. Under the first approach, filers and nonfilers are imputed by statistically matching tax units constructed in the CPS to tax returns in the IRS's Public Use File (PUF) on the basis of demographic characteristics and predicted income. (That method is used in CBO's individual income tax model.) In the second method, the linked data are used to derive the predicted probability of filing a federal tax return and the share of tax units that file; that information is then used to simulate filing in the CPS.

The two methods are evaluated on the basis of how similar the demographic and income characteristics of simulated filers and nonfilers are, in the aggregate, to those of filers and nonfilers in the linked data. Under both methods, the income distribution of simulated filers is broadly consistent with the income distribution of filers from the linked data. Simulated nonfilers using the predicted income method are less likely to have various sources of taxable income and receive lower income amounts, on average, than nonfilers in the linked data. That is due in part to the predicted income method classifying those with the lowest predicted income as nonfilers and also because nonfilers in the linked data potentially include individuals who file later or who are noncompliant taxpayers. Simulated nonfilers using the predicted probability of filing method also generally have lower income than nonfilers from the linked data, though they have higher income, on average, than simulated nonfilers under the predicted income method.

Filing Requirements

U.S. citizens, resident or abroad, and resident aliens are required to file a federal income tax return if their gross income over the tax year, which corresponds to a calendar year, exceeds the filing threshold or if they meet certain conditions.[5] Typically, the filing threshold is the total of the standard deduction, which varies on the basis of an individual's filing status and age, and the personal exemption.

A person's filing status is based on his or her marital status as of the last day of the tax year. The five categories are single, married filing jointly, married filing separately, head of household, and

[5] For tax purposes, resident aliens are noncitizens who either were lawful permanent residents of the United States at any time during the calendar year or were physically present in the United States on at least 31 days during the calendar year and 183 days during the current year and previous two years.

qualifying widow(er).[6] In addition to being unmarried, a head of household generally must pay more than half of the costs of maintaining the home in which he or she resides with a dependent relative or a qualifying child for over half of the tax year.[7]

A residency test also applies when claiming a child for certain tax benefits, such as the dependent exemption, the child tax credit, or the earned income tax credit (EITC). A qualifying child must live with the taxpayer for over half of the tax year, as well as meet certain conditions related to age and relationship to the taxpayer.

In some cases, the tax rules for filing status and dependents are not based on residency:

- A taxpayer can claim children and certain other relatives as dependents—regardless of where they reside—if he or she provides over half of the other person's support and that person has very low income (less than the amount of the personal exemption).
- A noncustodial parent can claim his or her children as dependents if the custodial parent has agreed as part of a child support arrangement. Even though the custodial parent does not receive the dependent exemption, however, he or she can still claim head-of-household filing status and the child-related EITC under certain circumstances.
- Unmarried taxpayers can claim head-of-household filing status if they claim their parent as a dependent and pay more than half of the costs for the home in which that parent lives—even if the taxpayer and parent reside in separate residences.
- Married taxpayers may file separately to avoid comingling their finances, even if they live together.

The filing thresholds for tax year 2006 were $8,450 for nonelderly single people, $16,900 for nonelderly married couples filing jointly, and $10,850 for nonelderly heads of households. Those thresholds were higher for people age 65 or older. For individuals who can be claimed as dependents by another taxpayer, the filing thresholds depend on whether income is earned or

[6] For more details, see Internal Revenue Service, *Exemptions, Standard Deduction, and Filing Information*, IRS Publication 501, https://go.usa.gov/xRsrr (209 KB).

[7] Qualifying widow(er)s must also pay more than half of the costs of maintaining the home in which they and a qualifying child reside, but a qualifying child must live with the taxpayer for the entire year. (Taxpayers are eligible for that filing status for two years after the death of their spouse.)

unearned and can be substantially lower than the filing thresholds for individuals who cannot be claimed as dependents. In 2006, individuals were also required to file if they owed the alternative minimum tax and other special taxes, received the advance earned income credit, had net earnings from self-employment exceeding $400, or earned wages of $108.28 or more from a church or qualified church-controlled organization that is exempt from payroll taxes.[8]

In other cases, individuals file even if it is not required. Such filers include individuals who had income tax withheld or who made estimated tax payments for the year. Some individuals may also qualify for refundable tax credits such as the EITC or additional child tax credit (the refundable portion of the child tax credit). In tax year 2001, about 87 percent of returns were filed because it was required, 11 percent of returns were filed to obtain a refund, and the remaining 2 percent were filed for no apparent reason.[9]

Generally, the filing deadline for a tax year is April 15th of the following calendar year. Taxpayers can receive a six-month extension by filing Form 4868. Taxpayers who reside outside of the United States and Puerto Rico receive an automatic two-month extension. Although most taxpayers file their returns in the following calendar year, about 3 percent of returns are filed late—typically in the subsequent two years.

Previous Research

Researchers have identified nonfilers by matching individuals from a nationally representative sample, using either survey data or other administrative data, to tax returns. Records from the population without a matching tax return are then identified as nonfilers. Linked data have also been used to examine how the same types of information (for example, reported income) are reported differently in separate data sources. More recently, some researchers have turned to tax administrative data to identify nonfilers.

[8] The Education, Jobs, and Medicaid Act of 2010 repealed the advance payment option for the EITC, effective for tax years beginning in 2011.

[9] See Janet Holtzblatt, "Trade-Offs Between Targeting and Simplicity," in James Alm, Jorge Martinez-Vazquez, and Mark Rider, eds., *The Challenges of Tax Reform in a Global Economy* (Springer, 2004), pp. 46–72.

One researcher, Jim Cilke, created a profile of nonfilers using linked CPS and 1040 data for the 1990 tax year and identified nonfilers as tax units in the CPS without a 1040 who were not required to file based on their income reported in the CPS.[10] At that time, both data sets could be directly matched to each other on the basis of Social Security numbers, and about 88 percent of individuals in the CPS had validated Social Security numbers that could be linked to tax returns and were included in his analysis. More than two-thirds of nonfilers in 1990 were composed of single dependents (typically students) and individuals age 62 or older without children. About 13 percent of the nonfiling population consisted of unmarried individuals with children. The presence of earned income, and to a lesser extent unearned income, increased the probability of filing. Individuals with higher adjusted gross income (AGI) relative to the filing threshold were also more likely to file.

Since 1990, several changes to the tax code have increased the incentives for people to file for a refund even if they do not meet the filing thresholds. First, the amount of the EITC was substantially increased, making it more advantageous to file a tax return. Second, eligibility for the EITC was extended to very low-income workers who do not live with children and who are at least 25 years old but under 65 years old. Third, the creation of new refundable tax credits—for purchases of homes, health insurance premiums, and tuition for higher education—increased the payoff from filing a tax return, regardless of whether individuals worked or had children. More recent research covers time periods that followed those expansions of refundable tax credits.

Another method of identifying nonfilers involves matching information returns filed by employers (Forms W-2) and other third parties (Forms 1099). Individuals with information returns but no 1040 are classified as nonfilers. Jacob Mortenson and his colleagues found that the size of the population estimated from tax returns and information returns closely approximated the U.S. resident population for tax year 2003.[11] About 30 million individuals were identified as

---

[10] See Jim Cilke, *A Profile of Non-Filers*, OTA Paper 78 (Office of Tax Analysis, July 1998), https://go.usa.gov/xRsrT (358 KB).

[11] See Jacob A. Mortenson and others, "Attaching the Left Tail: A New Profile of Income for Persons Who Do Not Appear on Federal Income Tax Returns," in *Proceedings: Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* (National Tax Association, November 2009).

nonfilers using that methodology. More than half of those nonfilers had Social Security income, which is typically received by elderly or disabled individuals, and almost 40 percent received labor income. Capital income and miscellaneous income such as rent, royalties, and crop insurance payments were much less common among the nonfiling population. Information returns were not available for certain types of income—most notably, earnings from self-employment. Researchers using tax data to analyze income inequality and intergenerational mobility have used information returns to represent nonfilers.[12]

Joshua Lawrence, Michael Udell, and Tiffany Young extended the analysis of information returns by using family structure characteristics from the CPS to create synthetic tax units and simulate the income tax liability of nonfilers.[13] For the 2005 tax year, they estimated that 38.6 million people in 22.8 million simulated tax units did not appear on a 1040, either as a taxpayer or a dependent. Most of those tax units were not required to file, even though in many cases they might have been able to receive a refund of overwithheld taxes or claim the EITC. (About 46 percent of simulated no-return tax units had wage income, though they might have been ineligible based on other unobserved characteristics.) About 40 percent of those tax units consisted of individuals age 62 or older who received Social Security income.

Examining how information for the same individual can vary across data sources is needed for undertaking tax modeling and understanding tax compliance. One measure of tax compliance, the voluntary filing rate, is the share of returns required to be filed that are filed on time. The IRS measures that rate using CPS data to estimate the denominator and tax return data for the numerator. To the extent that income reported on the CPS differs from what is reported on a 1040 or information return, the voluntary filing rate will be biased. Brian Erard, Mark Payne, and Alan

---

[12] See Jeff Larrimore, Jacob Mortenson, and David Splinter, *Household Incomes in Tax Data: Using Addresses to Move From Tax Unit to Household Income Distributions*, Finance and Economics Discussion Series 2017-002 (Board of Governors of the Federal Reserve System, January 2017), https://doi.org/10.17016/FEDS.2017.002; and Raj Chetty and others, "Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics*, vol. 129, no. 4 (November 2014), pp. 1553–1623, https://doi.org/10.1093/qje/qju022.

[13] See Joshua Lawrence, Michael Udell, and Tiffany Young, "The Income Tax Position of Persons Not Filing Returns for Tax Year 2005," in Alan Plumley, ed., *Recent Research on Tax Administration and Compliance: Selected Papers Given at the 2011 IRS-TPC Research Conference: New Perspectives on Tax Administration* (Internal Revenue Service, 2011), https://go.usa.gov/xRsrk.

Plumley found that some components of income were underreported in the CPS relative to information returns in tax year 2009.[14] Correcting for underreported income increases the estimated number of returns required to be filed and results in a smoother trend in the voluntary filing rate over time.

Data and Methods

For this project, I use a restricted-use data set created by the Census Bureau that links records in the 2007 CPS Annual Social and Economic Supplement (known as the March CPS) and federal individual tax returns filed in 2007 for the 2006 tax year to characterize filers and nonfilers and to model an individual's decision to file.

Current Population Survey

The March CPS is a nationally representative sample of the civilian noninstitutionalized population residing within the United States.[15] I treat the weighted CPS sample as representative of filers and nonfilers. Therefore, the characteristics of filers and nonfilers outside of the CPS sampling frame, such as people who are institutionalized or living outside the United States, are not considered. Although income data reported in the March 2007 CPS correspond to the amounts received in 2006, the demographic characteristics are as of the survey date in 2007.

To the extent possible using CPS data, I organize individuals into tax filing units based on their marital status and whether they can be claimed as a dependent by someone else in the household (based on age, relationship, and—in the case of adults—financial support). Spouses living in the same household are always part of the same tax unit, and those living apart are treated as either single or head of household. Within those constructed tax units, I refer to the householder (the

---

[14] See Brian Erard, Mark Payne, and Alan Plumley, "Advances in Nonfiling Measures," in Alan Plumley, ed., *New Research on Tax Administration: An IRS-TPC Conference, Papers Given at the 2012 IRS-Tax Policy Center Research Conference* (Internal Revenue Service, 2012), https://go.usa.gov/xRsb9. Another study compares the 2010 CPS to information returns reporting retirement income (1099-R forms) and finds that the CPS provides reasonable measures of the number of respondents with retirement income and the amounts of that income, although underreporting among the recipients with the largest amounts of pension income may lead to large differences between the two data sets in the average amount. See C. Adam Bee, *An Evaluation of Retirement Income in the CPS ASEC Using Form 1099-R Microdata* (U.S. Census Bureau, March 2013), https://go.usa.gov/xRsbX.

[15] The institutionalized population includes individuals residing in correctional institutions and nursing homes. Members of the military living off base or on base with civilian family members are included in the sampling frame of the CPS. Members of the military living on base in barracks, however, are not included.

individual who owns or rents the housing unit) and his or her spouse, respectively, as the primary and secondary taxpayer.[16] Their dependents are considered part of that tax unit. Because dependents can also file their own tax returns if they meet the filing requirements, each dependent can be the head of his or her own tax unit. For most of the analyses that follow, tax units headed by dependents are not included.

Complicated living arrangements present a challenge when constructing tax units. In some households, for example, a child might appear to be a dependent of more than one individual, but there is insufficient information in the survey to determine who can claim the child. That situation occurs in households that contain subfamilies who are related to the householder or spouse; those subfamilies consist of either a married couple without children or a parent (married or unmarried) with one or more never-married children under 18 years old. For example, a child who lives with a parent and grandparent in a residence owned by that grandparent could potentially be claimed as a dependent by either adult. The child and parent are considered a related subfamily. If the subfamily head can be a qualifying child or relative of the householder and does not have earned income, then the related subfamily becomes dependents of the householder's tax unit. (A subfamily that is unrelated to the householder is treated as its own tax unit.)

There are significantly more head-of-household filing units in the tax data than would be suggested by the information in the CPS. Additional adjustments are made to increase the number of constructed tax units who could file as head of household to be consistent with the number of tax returns with that filing status. Unlike other aspects of tax unit construction, which use observed information in the CPS and assume compliance with tax law, those adjustments are made to reflect the behavior of tax filers and may not conform to the legal criteria.[17] Because of the potentially large refund available to workers with children (through the EITC), some qualifying children are reallocated between unmarried partners and across related subfamilies

[16] If the residence is jointly owned or rented by a married couple, the householder can be either spouse. In those cases, the householder is treated as the primary taxpayer, and the spouse is the secondary taxpayer.

[17] For a discussion of how taxpayers modify their reported family status in response to EITC incentives, see Janet Holtzblatt and Janet McCubbin, "Issues Affecting Low-Income Filers," in Henry Aaron and Joel Slemrod, eds., *The Crisis in Tax Administration* (Brookings Institution Press, 2004), pp. 148–200.

with wages to create additional tax units that could claim the head-of-household filing status. However, a significant gap remains even after those adjustments.

I calculate income measures for the constructed tax units to be comparable to income reported on tax returns. Only income received by the primary taxpayer and spouse, if married, is included.[18] I calculate AGI based on the amounts reported on the CPS for wages and salaries, net self-employment income, unemployment compensation, retirement income, interest and dividends, rental income, alimony, survivors' benefits and disability income (except from workers' compensation), educational assistance, and the taxable portion of Social Security benefits. Several differences exist between AGI as reported on a tax return and AGI as calculated on the CPS. Most notably, the CPS does not include capital gains, which would be included in AGI on a 1040.

Federal Income Tax Returns

The IRS provided the Census Bureau with certain data from the Individual Master File (a data set containing tax forms for the entire population of tax filers) for individual income tax returns filed for tax year 2006, including amended returns for that year, as of October 2007. Although most tax returns for tax year 2006 were filed in 2007, those filed in later years are not included in that data set. To a limited extent, the data were edited by the IRS: Certain types of math and clerical errors were corrected, and Social Security numbers were verified.

The Internal Revenue Code and Treasury regulations limit the amount of tax return data provided to the Census Bureau.[19] The data provided contain information on the composition of the tax unit, such as filing status and number of exemptions by relationship to the taxpayer (spouse, child, parent, or other). The income data include amounts for some income sources (wages and salaries, dividends, interest, rent, and taxable Social Security), AGI, total income, and indicators

[18] In 2006, if a child was required to file a return and had unearned income below $8,500, a qualified parent could include the child's interest and dividend income on his or her return by filing Form 8814. About 0.1 percent of returns in 2006 used that option.
[19] See Internal Revenue Code, section 6103(j)(1); 26 C.F.R., part 301.

for the presence of most types of schedules, including those that report income from self-employment and investment.[20]

Linking the CPS to Tax Return Data

The Census Bureau creates a Protected Identification Key (PIK) that uniquely identifies individuals across data sets, making it possible to link records from various sources. PIKs are only available internally on restricted-use data sets; publicly available Census Bureau data sets do not include PIKs. The Census Bureau relies on a reference file (known as Numident data) that contains every Social Security number issued, along with names and dates of birth. Various federal administrative data provide other names and dates of birth, as well as addresses, associated with each Social Security number.[21] To protect the privacy and confidentiality of respondents, each Social Security number is randomly assigned a unique PIK, which the Census Bureau uses instead of the Social Security number.

For incoming records that already contain Social Security numbers (such as tax returns), the Social Security number, name, and date of birth are checked against the reference file. If there is a match, the corresponding PIK is assigned to that record. For records that do not contain Social Security numbers (such as records from the CPS) or records in which the Social Security number cannot be verified, the name, address, and date of birth from the incoming file are statistically matched to the reference file.[22] If there is a match, the record is assigned a PIK. If no statistical match can be found when the address is included, then a statistical match is performed using just the name and date of birth. Records that cannot be statistically matched to the reference file do not receive a PIK.

---

[20] The schedules include A (itemized deductions), C (profit or loss from business), D (capital gains or losses), E (supplemental income or loss), F (profit or loss from farming), and SE (self-employment). The Census Bureau's definition of total income differs from the IRS's by including nontaxable portions of pensions, annuities, and Social Security and by excluding taxable refunds, capital gains, and other gains.

[21] The administrative data include federal income tax returns, information returns (such as W-2s) filed with the IRS, Selective Service registrations, and administrative data from government programs including rental assistance, Medicare, and the Indian Health Service.

[22] Components of those variables are used to match records from both data sets. The records are deemed a match if the degree of agreement between the matching variables in the input and reference files exceeds a threshold value. For more details on the Census Bureau's methodology for assigning PIKs, see Deborah Wagner and Mary Layne, *The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software*, Working Paper CARRA WP 2014-01 (U.S. Census Bureau, July 1, 2014), https://go.usa.gov/xRsbm.

Information from various sources can be linked to the same individual as long as a PIK can be assigned in each data set. About 99 percent of tax returns available to the Census Bureau for tax year 2006 have a PIK for either the primary or secondary taxpayer (see **Table 1**).[23] Tax returns that do not have a PIK for the primary or secondary filer have lower AGI and more dependents, on average, than tax returns that have a PIK. Tax returns without a PIK are also more likely to be filed by individuals claiming the married filing separately status.

About 91 percent of constructed tax units in the CPS have a PIK for the primary or secondary taxpayer (see **Table 2**). Constructed tax units without PIKs have lower income, on average, though they are not more likely to receive means-tested transfer benefits. On average, constructed tax units without PIKs have fewer dependents than those with PIKs.

Reweighting. If the missing PIKs do not occur completely at random, estimates based on the subset of constructed tax units that contain PIKs will be biased. Survey records that contain errors in the name, address, or date of birth are unlikely to be assigned a PIK, though it is plausible that this occurs at random. Incomplete coverage and clerical errors in the Numident data can also prevent matches. An evaluation of the Census Bureau's PIK methodology found that rates of missing PIKs vary by region, socioeconomic status, and demographic characteristics.[24] It is problematic if the presence of a PIK is correlated with filing—for example, if individuals who work "off the books" are also less likely to file.

One way to account for the missing data and to minimize the bias of the estimates is to reweight the constructed tax units with PIKs by the inverse probability of having a PIK.[25] I estimate the

---

[23] A tax return may not be assigned a PIK if elements of the name and address on the tax return do not sufficiently match what is on the Numident. See Mary Layne, Deborah Wagner, and Cynthia Rothhaas, *Estimating Record Linkage False Match Rate for the Person Identification Validation System*, Working Paper CARRA-WP-2014-02 (U.S. Census Bureau, July 1, 2014), https://go.usa.gov/xR65b.

[24] For an evaluation of the Census Bureau's methodology of assigning PIKs, see Edward Mulrow and others, *Assessment of the U.S. Census Bureau's Person Identification Validation System* (submitted by NORC at the University of Chicago to the U.S. Census Bureau, March 31, 2011), http://tinyurl.com/y9v9eav7 (786 KB).

[25] See Bruce D. Meyer and Robert M. Goerge, *Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation*, Center for Economic Studies Discussion Papers CES 11-14 (U.S. Census Bureau, April 2011), https://go.usa.gov/xRsjd (126 KB).

likelihood that a constructed tax unit has any member with a PIK based on a set of demographic and income characteristics that have been found to be correlated with successful PIK assignment. I include characteristics of both the primary taxpayer (such as age and employment status) and of the tax unit (such as number of children and adults), as well as household income as a percentage of the federal poverty level (see **Table 3**). I also include a variable for whether the Census Bureau imputed all of the responses for the primary taxpayer because he or she did not answer any of the questions on the survey. I calculate the predicted probability of receiving a PIK for each constructed tax unit using the estimated coefficients. I multiply each tax unit's weight (the CPS household weight) by the inverse of the predicted probability of receiving a PIK and use adjusted weights in the analysis of the linked data. Reweighting reduces the bias of the estimates but does not completely eliminate it because there are likely to be unobservable characteristics associated with successfully getting a PIK.

Linked Data. The constructed tax units with PIKs from the CPS form the population for the analysis of the characteristics of filers and nonfilers and for modeling the probability of filing a tax return. Within the CPS, I construct 93,000 nondependent tax units, which—using the CPS household weights—represent 292.7 million individuals in 141.7 million tax units (see **Table 4**). About 7,900 (or 9 percent) of those constructed tax units do not contain any member with a PIK and are dropped from this analysis because they cannot be linked to any 1040. After reweighting, the remaining constructed tax units also represent about 141.7 million nondependent tax units.

Of those, I classify 117.9 million tax units as filers because at least one member of the constructed tax unit is the primary or secondary taxpayer on a 1040. (Those constructed tax units are associated with 126.2 million tax returns because some tax units are linked to more than one return.) I classify the remaining 23.7 million constructed tax units as nonfilers. In most of those cases, no one in the constructed tax unit can be matched to a 1040, though in a small share of cases a member of the constructed tax unit is claimed as a dependent on a 1040. The linked data (henceforth referring only to the reweighted sample of constructed tax units in which a member has a PIK) consist of 251.9 million individuals in constructed tax units that file and 40.7 million individuals in tax units that did not file.

Characteristics of Filers and Nonfilers Using Linked Data

I examine the income and demographic characteristics of constructed tax units in the linked data by the presence of a 1040. I refer to constructed tax units in which a member is a primary or secondary taxpayer on a 1040 as filers and unmatched nondependent units as nonfilers. In some cases, a constructed tax unit matches more than one 1040, or the number of dependent exemptions claimed on the 1040 differs from the number of people in the constructed tax unit. Those tax units are counted as filers, though I explore those discrepancies in the **Appendix** to this paper. To enable comparisons between filers and nonfilers, I present the characteristics reported in the CPS. Constructed tax units are categorized by whether the primary or secondary taxpayer is age 65 or older; the remaining units are grouped by marital status and household composition: married with dependents, married with no dependents, unmarried with dependents, and unmarried with no dependents.

Constructed Tax Units With 1040s

Of the nearly 118 million constructed tax units linked to a nondependent tax return, half are nonelderly and do not have dependents (see **Table 5**). Nonelderly unmarried taxpayers without dependents account for the largest share of filers—33 percent. About a quarter of filers are nonelderly married couples with dependents. Only 11 percent of filers are nonelderly unmarried filers with dependents. About 15 percent of tax units that file contain at least one taxpayer age 65 or older.

Tax units that file a tax return typically receive income from some taxable source, according to information reported in the CPS. The most common income source—wages and salaries—is received by 82 percent of all filers (see **Table 6**, top panel). Interest income is the most common source of unearned income, though receipt across the demographic groups varies from about 27 percent for unmarried filers with dependents to 64 percent for filers age 65 or older. Filers age 65 or older are more likely to have unearned income than earned income. About 27 percent of those filers have income from wages and salaries, but 91 percent receive Social Security benefits, and nearly two-thirds reported interest income.

Overall, a much smaller share of filers report participating in nontaxable transfer programs than report wages and salaries. Medicare and Medicaid are the most common benefits received. Unmarried filers with dependents are more likely to receive benefits through the Supplemental Nutrition Assistance Program (SNAP) or Medicaid than are other filers.

On average, income from wages and salaries accounts for the largest share of AGI for filers (see **Table 6**, bottom panel). Nonelderly married filers have higher AGI, on average, than other tax unit types. Among nonelderly tax units, only a small share of AGI is derived from unearned income. In contrast, average amounts of unearned income (from dividends, interest, rent, and Social Security benefits) are higher for tax units with members age 65 or older than for other filers.

Constructed Tax Units Without 1040s

Compared with filers, nonfilers in the linked data are less likely to consist of nonelderly married couples and more likely to contain someone who is age 65 or older or who is unmarried (see Table 5). About 31 percent of nonfiling tax units contain individuals age 65 or older, and 37 percent consist of nonelderly unmarried individuals without dependents. Although constructed tax units with married couples account for 42 percent of tax units that file, they account for only 17 percent of those that do not file.

Nonfilers are less likely than filers to receive income from taxable sources (**see Table 7**). Wages and salaries are the most common source of income for nonelderly nonfilers, though receipt ranges from 47 percent among unmarried tax units with dependents to 80 percent of married tax units with dependents. Unearned taxable income (such as dividends and interest income) is also less common among nonfilers than filers. Less than one-quarter of nonfilers report interest income, and fewer than 10 percent report dividends or rental income. Compared with filers, a greater share of nonelderly nonfilers receive income from Social Security (which includes disability insurance payments).

Receipt of means-tested assistance is more common among nonfilers than filers. About 17 percent of nonfilers receive SNAP benefits, compared with about 4 percent of filers. Fourteen percent of nonelderly units without 1040s receive Medicare, but only 5 percent of filers under the age of 65 receive that benefit. Medicaid benefits were received by 29 percent of all nonfilers and 11 percent of all filers.

Overall, nonfilers have lower average income amounts from taxable sources than do filers. On average, nonfilers have AGI of $20,000. Average wages and salaries among married nonfilers range between about $34,000 and $52,000, and average self-employment income among nonfilers is comparable to that of filers. In contrast to filers, nonfilers receive higher amounts of Social Security benefits and means-tested assistance, on average.

Comparing the Linked Data With Other Tax Return Data

Both the CPS data and the tax information provided by the IRS to the Census Bureau have limitations that affect the construction of the linked data and the number of filers and nonfilers derived from that data. First, the CPS does not include individuals residing outside the United States or in institutional settings (for example, nursing homes and prisons). Thus, the linked file is limited to U.S. residents living in noninstitutional settings. Second, the tax data do not include 2006 returns filed after October 2007. As a result, late filers are classified as nonfilers in the linked file.

To compare the linked data with a comprehensive sample of tax returns, I use information from the Internal Revenue Service's Public Use File. The PUF is a stratified random sample of individual income tax returns that is weighted to be representative of the tax returns filed for a tax year. To protect the identity of taxpayers, the IRS modifies information in various fields so that the resulting records do not contain complete information from any individual tax return. Unlike the linked data, the PUF includes tax returns filed by taxpayers living abroad, and those returns generally can be distinguished from returns filed by taxpayers residing in the United States. It also contains tax returns that were filed in 2007 but cover the three previous tax years. Other differences in the number of returns between the PUF and the linked data are not as easily

observed. For example, some returns would be filed by people who are not included in the CPS sampling frame—individuals who were institutionalized in March 2007, died before the survey date, or were members of the military living in barracks.

Comparing Tax Returns in the Linked Data and the PUF

I compare the characteristics of tax filers in the PUF to the constructed tax units with 1040s in the linked data after making two adjustments to the PUF to make it comparable to the linked data. First, I remove tax returns that suggest the taxpayer lives abroad (based on having an address outside of the United States, paying the foreign earned income tax, or claiming the foreign tax credit) from the PUF.[26] Next, I remove late returns. Because taxpayers who file their 2006 returns after October 2007 are initially classified as nonfilers in the linked data, the number of nonfilers in the linked data set is too high. After removing tax returns in the PUF that potentially belong to taxpayers residing abroad or that were received by the IRS in 2007 but were filed for tax years 2003–2005, 117.5 million tax returns are left (see **Table 8**, top panel). That number is similar to the number of filers (117.9 million) in the linked data.

Compared with the distribution of tax returns in the PUF (after removing foreign returns and late filers), the filers in the linked data are less likely to be unmarried with dependents and more likely to be married. The income amounts reported in the PUF are generally very similar to those reported in the CPS by filers in the linked data. Because not all components of AGI are reported on the CPS, the share of filers in the linked data with nonzero AGI (96 percent) is slightly lower than the share of tax returns in the PUF with nonzero AGI (99 percent). A slightly higher share of returns in the PUF report wage and salary income (84 percent) compared with filers in the linked data (82 percent). Filers report wages and salaries totaling $5.6 trillion in the CPS, compared with the $5.5 trillion of wage and salary income reported in the PUF.

[26] Those three different variables were used to determine U.S. residency, but each had limitations. The state identifier in the PUF (which also indicates foreign residence) is not inclusive of all taxpayers who file from abroad because it is missing on records with very high or very low AGI. Additional taxpayers who could be abroad were identified as those with foreign earned income tax or those who claim a foreign tax credit. That may overestimate the number of taxpayers living abroad because taxpayers living in the United States and paying foreign taxes, such as on dividends from foreign corporations, can also claim the foreign tax credit. However, those variables do not identify all taxpayers residing abroad either. For example, income from the U.S. government for employees stationed abroad is not considered foreign earned income, and that income may not be taxed by the host country.

Accounting for Late Filers in the Linked Data

Taxpayers who file after October 2007 appear to be nonfilers in the linked data, but they may differ from people who never file. To illustrate the effect of late filers in the sample of nonfilers, I adjust the identification of nonfilers in the linked data using information about late filers in the PUF. First, I assume that the share of returns in the 2006 PUF that are filed for the 2003–2005 tax years—about 3 percent—is the same as the share for the 2006 tax year that will be filed late. Then I categorize nonfilers in the linked file and late filers in the PUF by filing status and wage bin (under $20,000, $20,000 to $49,999, $50,000 to $99,999, $100,000 to $199,999, and $200,000 and over) and calculate the ratio of late filers in the PUF to nonfilers in the linked data for each group. For each nonfiler in the linked data, I draw a random number; if it is less than that ratio, I reclassify the unit from a nonfiler to a filer. That method reduces the estimate of nonfiling units in the linked file from 23.7 million to 20.8 million (see **Table 8**, bottom panel).[27]

Even after removing potential late filers from the pool of nonfilers, total income from taxable sources for nonfilers identified through the linked data is generally higher than that for nonfilers identified through a match between information returns and tax returns. Based on the amounts reported in the CPS, about 40 percent of nonfilers have wage and salary income totaling $347 billion. Reclassifying some nonfilers as late filers reduces the share with wages to 36 percent and total wage income to $236 billion (see **Table 9**). In comparison, researchers matching information returns to tax returns estimate that about one-quarter to one-third of nonfiling individuals have wage income.[28] Several factors may account for that discrepancy—an incomplete match between the 1040s and CPS would result in too many nonfilers (which the reweighting does not adequately correct); some employers pay their workers "under the table" and do not file W-2s; or respondents may overstate their earnings in the CPS. Because of

---

[27] That imputation method switches fewer tax units (about 3 million) from nonfilers to filers than the implied number of late returns in the PUF (3.5 million). One reason for that difference may be that some constructed tax units in the linked data may be associated with multiple tax returns.

[28] For more details, see James Cilke, "The Case of Missing Strangers: What We Know and Don't Know About Nonfilers" (paper presented at the 107th Annual Conference of the National Tax Association, Santa Fe, New Mexico, November 13–15, 2014), http://tinyurl.com/y7chw254 (403 KB); and Joshua Lawrence, Michael Udell, and Tiffany Young, "The Income Tax Position of Persons Not Filing Returns for Tax Year 2005," in Alan Plumley, ed., *Recent Research on Tax Administration and Compliance: Selected Papers Given at the 2011 IRS-TPC Research Conference: New Perspectives on Tax Administration* (Internal Revenue Service, 2011), https://go.usa.gov/xRsrk.

uncertainty over how to best identify late filers among the nonfilers, I do not adjust the sample of nonfilers for late filers in the main analysis.

Identifying Nonfilers Through Statistical Matches

The linked microdata are generally unavailable to researchers outside of the Census Bureau, so alternative methods of simulating nonfilers using publicly available data are still necessary. I describe two methods of statistically matching tax returns from the PUF to CPS records—one using predicted income and the other using the predicted probability of filing. (For purposes of this analysis, all returns from the PUF and all records from the CPS are used.)[29] Characteristics of simulated filers and nonfilers from those two methods are then compared with the characteristics of filers and nonfilers in the linked data set.

Predicted Income

Under this method, tax returns from the PUF are statistically matched to CPS records with similar demographic attributes on the basis of predicted income. CBO uses this method to model individual income taxes.

- First, tax returns are divided into groups based on marital status, the number of taxpayers age 65 or older, the number of dependents in the tax unit (0, 1, or 2 or more), and whether the return is filed by a dependent. Because the PUF does not contain age, taxpayers over age 65 are identified by their use of the higher standard deduction available to them or by the presence of certain forms of income (such as Social Security or pension income).
- Second, an aggregate income measure is created to get comparable income measures on both the tax return and the constructed tax unit. Within each demographic group, total income (gross income net of taxable Social Security income) is regressed on several income sources (which are also found in the CPS) using income values from the PUF. Predicted income is calculated for each tax return.

Next, the CPS records are arranged in a comparable way.

[29] Notably, neither the PUF nor the publicly available CPS contain PIKs.

- CPS individuals are grouped into tax units (as described in an earlier section), and constructed tax units are categorized into the same demographic groupings as the tax returns.
- Within each demographic group, predicted total income is calculated for each constructed tax unit by applying the estimated coefficients from the PUF to the amounts reported in the CPS for the comparable sources.

Modeling Filers and Nonfilers. Because of the differences in sample weights between the two files, the match is not performed on a one-to-one basis. Typically, within each demographic group, the match starts with the record from each file with the highest predicted total income.[30] Of the two records, the one with the lower sampling weight is matched to only one corresponding record from the other file. The record with the higher weight is "split," and it is available (with its weight reduced) to be matched to the next record in the other file. Because there are more constructed tax units in the CPS than tax returns in the PUF, the unmatched tax units will be classified as nonfilers.

By design, this method matches every tax return in the PUF to a constructed tax unit. Doing so has advantages for analysts that are using the tax filing population to model tax revenues because the tax records reflect households that remitted tax in that year. The number of nonfilers is the difference between the number of constructed tax units and the number of tax returns in the PUF and probably represents a lower bound because some returns in the PUF are filed by taxpayers who are not included in the CPS. Without any adjustments, the number of constructed tax units in the CPS exceeds the number of nondependent filed returns in the PUF by 13.6 million.

Comparison With Linked Data. Under this method, each tax return in the PUF is assigned to a constructed tax unit in the CPS with the same demographic features. As a result, the number of simulated filers in the CPS and their demographic composition match those in the PUF. This method simulates about 10 million more filers than are in the linked data and, conversely, 10

[30] Some records were matched across demographic groups when the weighted number of SOI returns in a subgroup exceeded the weighted number of constructed tax units for that filer type in the CPS. That mostly affected tax returns claiming head-of-household status, who were then matched to constructed tax units headed by single filers.

million fewer nonfilers (see **Table 10**). A larger share of simulated filers is both unmarried and under the age of 65 compared with filers in the linked data.

Overall, the share of simulated filers with various income sources is within 2 percentage points of the share of filers with income (see **Table 11**). Relative to filers in the linked data, simulated filers who are age 65 or older are generally more likely to have income from taxable sources, and simulated filers who are younger and unmarried, especially those with dependents, are less likely to have income from various taxable sources. The largest differences occur in wage and salary income, where the share of unmarried simulated filers with earnings is between 6 and 9 percentage points lower than in the linked data. A larger share of simulated filers who are unmarried—between 1 and 5 percentage points—report means-tested transfer benefits compared with filers in the linked data.

The average income amounts from taxable sources for simulated filers are within 6 percent of the averages for filers in the linked data. Simulated filers who are unmarried tend to have lower average income amounts than unmarried filers in the linked data. Other simulated filers tend to have higher income amounts, on average, than comparable filers in the linked data.

The differences between nonfilers based on the linked data and simulated nonfilers are more pronounced (see **Table 12**). Simulated nonfilers are those with the lowest predicted income in their respective demographic bins, so smaller shares of each group receive the various sources of taxable income, and average income amounts are substantially lower than those for nonfilers in the linked data. Among simulated nonfilers who are unmarried, average income amounts from taxable sources are close to zero.

To test the accuracy of the predicted income approach, I compared how the simulated results using publicly available data compared with the observed filing behavior in the linked data. Overall, the predicted income method simulates the same filing behavior as is observed in the linked data for 83 percent of tax units—though it is markedly better at projecting filers than nonfilers. About 94 percent of filers in the linked data are simulated as filers, but only about 27 percent of nonfilers in the linked data are simulated as nonfilers. Nonelderly filers who are

correctly classified by the model have similar AGI, on average, in the CPS and on the 1040 (see **Table 13**, top panel), even though the AGI calculated on the basis of the CPS does not contain information about expenditures that can be deducted from AGI or some components of taxable income, notably capital gains and losses. Average AGI for filers age 65 or older in the CPS is about half the amount on the 1040, which may be due to unearned income accounting for a larger portion of income for older taxpayers. Reported wages and salaries on the CPS are slightly higher than on the 1040, though that might be due to the exclusion of pretax contributions to retirement accounts from wages and salaries on the 1040.

The misclassified nonfilers—filers in the linked data who are simulated as nonfilers—predominantly are age 65 or older or are nonelderly married individuals (see **Table 13**, bottom panel). On average, the misclassified nonfilers report substantially lower income amounts on the CPS than on the 1040. Among that group, about two-thirds of simulated nonfilers who are nonelderly and married, and fewer than 5 percent of other simulated nonfilers, report wage and salary income in the CPS (not shown in Table 13). Average wages and salaries reported in the CPS by each group was under $11,000. In contrast, on the basis of the 1040s, about 83 percent of misclassified nonelderly individuals reported wage income, and average wages ranged between about $2,000 and $20,000.

Predicted Probability of Filing

An alternative way of ranking constructed tax units within each demographic group for the statistical match uses the likelihood that a constructed tax unit files a return instead of predicted income. The higher the predicted probability of filing, the more likely a constructed tax unit will be matched to a tax return.

Modeling the Filing Decision. I model the likelihood that a constructed tax unit files a return based on observed characteristics from the CPS using the sample of constructed tax units with PIKs from the linked data. The dependent variable, whether a constructed tax unit files, is the presence of a 1040 in that linked data. I define a constructed tax unit headed by a nondependent as a filer if the head or spouse is a primary or secondary taxpayer on a 1040. A constructed tax unit headed by a dependent taxpayer is defined as a filer if he or she is the primary taxpayer on a

1040 and does not claim the personal exemption.[31] Similar to the set of covariates used by Cilke, the regressors include the log of gross income (excluding taxable Social Security income) that would be reported on a tax return, an indicator for negative gross income, receipt of a means-tested transfer (SNAP, housing assistance, Temporary Assistance for Needy Families, or energy assistance), the number of household members who receive Medicaid, indicators for various income sources (interest, dividends, self-employment, rent, wages and salaries, retirement, and Social Security), and indicators for amount of education and race. The models are estimated separately for each demographic group based on marital status, number of dependents, whether the filer is age 65 or older, and whether the taxpayer can be claimed as a dependent.

Filers are simulated on the basis of the unit's predicted probability of filing so that the share of simulated filers matches the share of filers in the linked data. I rank constructed tax units in each group by the predicted probability of filing, and the tax units with the lowest predicted probabilities (to match the share of nonfilers from the linked data in each group) are simulated as nonfilers. The number of simulated filers using that method can differ from the number of tax returns in the PUF, even after removing returns filed for prior tax years and from abroad, because it will depend on the number of constructed tax units in the CPS and the estimated share without a linked 1040. The number of simulated filers using that method can be thought of as the number of civilian noninstitutionalized filers who file on time.[32]

Using a probit model, I estimate the probability that a constructed tax unit has a 1040 for each demographic group (see Table 14). Filing rates across groups of nondependent taxpayers range from about 60 percent among tax units headed by unmarried individuals age 65 or older to more than 90 percent among units headed by married nonelderly couples.[33] About 10 percent of nonelderly dependents file their own returns. Individuals with higher gross income (total income net of Social Security income) and those with income from taxable sources are more likely to

[31] That definition excludes dependents who claim the personal exemption as filers.
[32] Alternatively, the statistical match can incorporate randomness by assigning a random number to each constructed tax unit and classifying them as a filer if that random number is less than the tax unit's predicted probability of filing.
[33] Removing late filers from the count of nonfilers and including them as filers reduces the share of nonfilers in each group by between 1 and 3 percentage points and results in similar estimated coefficients.

file. The presence of wage and salary income, in particular, is strongly associated with filing a tax return. That probably is because workers have income taxes withheld or qualify for the EITC if they have sufficient taxable income. The receipt of means-tested transfers or Medicaid generally is negatively associated with filing.

Comparison With Linked Data. By construction, the share of simulated filers in each demographic group using the predicted probability method matches that in the linked data (see Table 10). The predicted probability method finds slightly fewer filers than are in the linked data.

The share of simulated filers with various sources of income is typically higher than the corresponding share of filers in the linked data with income from that source. In particular, simulated filers are slightly more likely to report wage and salary income than are filers (see **Table 15**). Across demographic groups, simulated filers have higher AGI and wages, on average, than filers, though the differences are within $4,000.

Larger differences appear between nonfilers based on the linked data and simulated nonfilers than between filers in the linked data and simulated filers. Except for self-employment income, simulated nonfilers using the predicted probability of filing method are generally less likely to have income from various taxable sources than nonfilers in the linked data (see **Table 16**). Across all demographic groups, the share of simulated nonfilers with wage and salary income is about 22 percentage points lower than the share of nonfilers with wages. Larger shares of simulated nonfilers receive means-tested benefits than do nonfilers. Average incomes from taxable sources for simulated nonfilers are generally lower than those for nonfilers. On average, the AGI of simulated nonfilers under the predicted probability method is about one-third of the average AGI of nonfilers in the linked data. Average wage and salary income of simulated nonfilers is only about one-fifth of the average wage and salary income of nonfilers in the linked data.

Overall, the simulated filing behavior from the predicted probability method matches the outcome from the linked data for 84 percent of constructed tax units. Compared with the predicted income approach, using the predicted probability approach means that a slightly lower

share of filers in the linked data—91 percent—are simulated as filers, but a higher share of nonfilers in the linked data—53 percent—are correctly simulated as nonfilers. Except for filers who are age 65 or older, filers who are correctly simulated as filers through the predicted probability method have AGIs on the CPS and 1040 that are, on average, within $6,000 of each other (see **Table 17**). The discrepancy in average AGI for filers who are age 65 or older may stem from underreporting of retirement income and unearned income on the CPS, which accounts for a larger share of income for older taxpayers. Differences in average earnings are much smaller across older filers than among those under the age of 65. Filers who are simulated as nonfilers under the predicted probability method, on average, also report lower income amounts in the CPS than on the 1040.

How the Alternative Simulation Methods Compare

The two simulation methods differ in determining which constructed tax units from the CPS are more likely to be filers and the total number of simulated filers. The predicted income method matches all tax returns in the PUF to constructed tax units in the CPS using only demographic information and predicted income. Because the PUF contains more tax returns than the linked data, that method simulates more filers than appear in the linked data. In contrast, applying the predicted probability of filing method to the CPS results in the same share of units that are simulated filers as in the linked data. The number of simulated filers from that method differs from the number of simulated filers derived from the predicted income method because the number of filers in the linked data does not match that in the PUF, even after removing tax returns that are not in the sampling frame of the CPS. Unlike the predicted income method, the predicted probability method uses additional demographic information and receipt of nontaxable income sources to model filing behavior. Although the predicted income method probably results in too few simulated nonfilers, the predicted probability method may simulate too many nonfilers by treating late filers as nonfilers.

Simulated Filers. Under the two methods, the overall shares of simulated filers who receive income from taxable sources are within 6 percentage points of each other. The largest discrepancy occurs for simulated filers who are unmarried: Using the predicted income method, the share with wage and salary income is between 12 and 18 percentage points lower than when

using the predicted probability method. Average income amounts from taxable sources for all simulated filers are within 8 percent of each other. Unmarried individuals that are simulated as filers under the predicted income method generally have lower income, on average, than simulated filers under the predicted probability method. In contrast, the average income of filers who are either elderly or married is higher in the simulations using the predicted income method.

Simulated Nonfilers. Total income of simulated nonfilers under both methods is lower than that of nonfilers in the linked data, even after removing potentially late filers. The simulated nonfilers under the predicted probability method have about $26 billion more in wage and salary income than do simulated nonfilers from the simulated income method, and both methods result in totals that are substantially lower than the $247 billion in wage and salary income of nonfilers after removing late filers. Simulated nonfilers are less likely to have various types of taxable income and have lower average income relative to nonfilers. Like nonfilers in the linked data, most simulated nonfilers have wages and salaries below $5,000, but neither simulation method is able to match the distribution of nonfilers with higher wages and salaries. Other than for married taxpayers without dependents, the predicted probability method simulates the sources and average amounts of income of nonfilers that more closely approximate those characteristics of nonfilers than does the predicted income method.

Conclusion

This analysis uses the linked CPS and tax return data to create a descriptive profile of filers and nonfilers and to evaluate how closely the income and demographic characteristics of simulated filers and nonfilers correspond to those of actual filers and nonfilers. The two simulation methods considered here statistically match records from the PUF to the CPS based on the rank order of records from each source. The method used in CBO's individual income tax model ranks tax units within each demographic group by predicted income. An alternative method incorporates the predicted probability of filing derived from the linked data.

The share of constructed tax units that is correctly simulated (that is, a simulated filer has a 1040 in the linked data or a simulated nonfiler does not have a 1040) is similar under both methods. By construction, the demographic characteristics of simulated filers using the predicted

probability method match those of the linked data. That construction was not applied to the predicted income method, and there were relatively more unmarried individuals without dependents among the simulated filers than is observed in the linked data. The share of simulated filers under either method receiving various income sources is generally similar to that of filers in the linked data. The differences in average AGI and wage and salary amounts between simulated filers and filers in the linked data are within $2,000, with smaller differences for income from other sources. The predicted income approach simulates filers who are unmarried with income amounts that are, on average, lower than those of filers in the linked data, but the predicted probability method simulates filers who have higher income amounts, on average, than actual filers.

Simulated nonfilers are generally less likely to have income from various sources and have lower average income amounts than nonfilers in the linked data, though the differences in average income are larger under the predicted income method than under the predicted probability method. Under both simulation methods, constructed tax units with higher incomes are more likely to be classified as filers than as nonfilers. Although the predicted probability method uses additional covariates, total income and measures correlated with income (such as the receipt of means-tested benefits) are strong predictors for filing.

The results have several implications for the statistical match used in CBO's tax model. First, nonfilers in the linked data have income that is generally higher than the income of simulated nonfilers using the predicted income method from CBO's tax model. Although the income characteristics of simulated nonfilers using predicted probability are closer to those of nonfilers, additional analysis would be needed before incorporating the predicted probability ranking into CBO's tax model. Linked data from additional years would be needed to confirm that this method is an improvement over the predicted income method for other years as well, because the tax model is updated periodically as more recent tax return data become available. Second, both of the methods considered here rank records deterministically, but tax units near the simulated filer cutoff may resemble both filers and nonfilers. Investigating whether incorporating randomness into the ranking algorithm would result in simulated nonfilers that more closely resemble nonfilers is an area for future work.

Finally, analysis of the linked data suggests that the predicted income method simulates too few nonfilers. The number of simulated nonfilers would increase if fewer tax returns in the PUF were used in the match or if more constructed tax units were created in the CPS. One refinement to the matching methodology would be to identify and remove tax returns in the PUF that are filed by tax units outside of the CPS sampling frame (for example, by people residing abroad). Further analysis would focus on constructing tax units in the CPS that better reflect the tax units found on the 1040. In some cases, constructed tax units based on the relationships reported in the CPS do not align with dependents claimed on the 1040. Because the composition of the tax unit affects an individual's filing status, whether the filing threshold is met, and potential tax liability, accurately assigning individuals into tax units would be useful for any statistical match.

Identifying nonfilers as survey respondents without a matching 1040 in the linked data has advantages over relying on tax or survey data alone, though it may still be incomplete. Not every survey respondent has the unique identifier that enables a match; a very small number of tax returns lack that identifier as well. Furthermore, I do not observe tax returns for tax year 2006 that are filed after 2007, so some tax units that eventually file will appear to be nonfilers. Additional work to examine how late filers differ from taxpayers who file on time may be useful in modeling nonfilers. Compared with other approaches that use information returns to identify nonfilers, methods using the linked data have both advantages and disadvantages. The data on taxable income reported in information returns are generally more complete than the data in the linked file; however, the linked data contain more information on self-employment and nontaxable income. In addition, the linked file contains information on family characteristics that is not available from the information returns.

Appendix: How Constructed Tax Units Differ From Tax Units on 1040s

A contributing factor to differences between simulated nonfilers and nonfilers based on linked data are differences in how comparable information is reported in the Current Population Survey (CPS) and on 1040s. That information can differ for several reasons. The CPS contains demographic characteristics as of the survey date in 2007, but the filing status on a tax return is based on a taxpayer's marital status as of December 31, 2006, and the number of dependents can depend on whether a taxpayer and child lived together for at least six months during 2006. Taxpayers who marry or divorce or whose dependent moves in or out between the end of 2006 and the CPS survey date in 2007 will be classified differently in the two data sets. A CPS household includes only individuals who reside together, but tax rules for filing status and dependents are not always based on residency—which can lead to a constructed tax unit bearing little resemblance to the tax unit reported on the 1040.

The two data sets also differ in the sources and amounts of income they report. On the 1040, taxpayers can refer to information returns filed by the payer, but income data on the CPS may be more prone to recall error. The amounts reported can also differ because the 1040 focuses on the taxable amounts (for example, wages and salaries do not include pretax contributions to retirement accounts), whereas a survey respondent is asked about earnings before any deductions. Some income sources, such as capital gains, are included in adjusted gross income on a 1040 but are not measured in the CPS. In addition, the Census Bureau replaces missing survey responses with imputed values that draw on other information from a respondent's survey or from comparable respondents. Imputations, particularly those that draw on another survey participant's responses, can also contribute to differences in income reported in the two data sets.

Misreporting on tax returns will also result in differences between the constructed tax units and the filing units on tax returns. In some cases, individuals may intentionally misreport their income, filing status, or number of dependents on their tax return to reduce their tax liability. In other cases, taxpayers may make unintentional errors because they do not understand complicated tax rules.

The linked data set provides some insight into how often the constructed tax units based on CPS data correspond to the filing units on 1040s—as well as some of the reasons for those differences.

Membership in the Tax Unit

I examine the correspondence between constructed tax units and tax units as they appear on a tax return using the linked data. Because I link CPS records and tax returns on the individual level, a constructed tax unit can be linked to multiple tax returns or to a return that includes different people in the tax unit. To evaluate whether the constructed tax unit is the same as the tax unit on the 1040, I compare the number of people in the constructed tax unit to the number of personal exemptions on the linked 1040s.[34]

In about 101 million (or 71 percent) of constructed tax units, all members of the tax unit can be linked to the same 1040 (see **Appendix Table 1**). About 9 million (or 7 percent) of constructed tax units match to one 1040, but not everyone in the tax unit can be linked to it. That situation can arise because a member of the constructed tax unit is missing a Protected Identification Key (PIK) and cannot be linked to any 1040, the 1040 does not have PIKs for all dependents, or some members of the constructed tax unit are not claimed as dependents. A small share of constructed tax units matches to more than one nondependent 1040. That can occur in several ways—a married couple files two returns separately, a dependent filer incorrectly claims the personal exemption, or the constructed tax unit erroneously includes individuals who cannot be claimed as dependents. Nearly 24 million (or 17 percent) of constructed tax units cannot be matched to a 1040.

Comparing the number of people in a constructed tax unit to the number of exemptions on the 1040s provides insight regarding the reasons that not all members can be linked to a 1040—in particular, whether the constructed tax unit contains multiple tax units or whether the discrepancy stems from an inability to match caused by missing PIKs. For most constructed tax

[34] I compare tax units on the basis of unit size instead of examining matches using the PIKs of the dependents because tax returns only contain PIKs for up to four dependents. As a result, not all dependents in larger constructed tax units can be matched to a tax return. Children are also slightly less likely to receive a PIK than adults because the reference files generally contain less information about children.

units, there is only one linked 1040, everyone in the constructed tax unit is linked to it, and the number of people in the constructed tax unit is the same as the number of personal and dependent exemptions. When only some members of the constructed tax unit can be linked to a 1040, the total number of exemptions on the 1040 is usually smaller than the size of the constructed tax unit—which suggests that those unlinked individuals in the constructed tax unit should be in a separate tax unit because they are not claimed as dependents. Furthermore, the majority of those unlinked individuals have a PIK, which would enable a match if they were claimed as dependents on the tax return. Some of the partial matches can be attributed to the reallocation of children across taxpayers to increase the number of constructed tax units with head-of-household filing status. About 11 percent of constructed tax units in which all members match to the same 1040 are head-of-household filers, compared with about 43 percent of partially matched units. However, constructing tax units without reallocating children across unmarried partners does not substantially change those results, because a relatively small number of tax units are affected by reallocating children. The number of dependents claimed on the 1040s exceeds the number of dependents in the CPS by about 11 million. That discrepancy suggests that there may be individuals who are not in the CPS household who were part of the tax filing unit.

A small share of constructed tax units matches to multiple 1040s, which can reflect multiple tax units in the family (for example, if some members of a tax unit cannot be claimed as dependents) or changes in family composition (for example, a couple in the CPS who married in early 2007 would have filed two separate returns for the 2006 tax year). About 12 percent of constructed tax units linked to multiple returns consist of married taxpayers filing separately. In most cases, though, the multiple 1040s that match to those constructed tax units have filing statuses that suggest constructed tax units should be split into multiple units to reflect erroneous filing behavior. In 30 percent of those cases, one spouse files as single and the other files jointly, while in another 23 percent of cases one spouse files as single and the other as the head of household.

Filing Status

Systematic differences in reported income and demographic characteristics in the CPS and 1040 affect the statistical match of constructed tax units to tax returns. Both are classified on the basis

of their demographic characteristics—marital status, age, and presence of dependents—so they can be statistically matched to similar units. Errors can arise, if, for example, a constructed tax unit with a married couple with two children actually files two tax returns as heads of households. In the statistical match, that constructed tax unit would be matched to a tax return with the married filing status, and there would be a shortfall in the number of constructed tax units that consist of unmarried parents with one child that can be used to match to the tax returns of heads of households.

Eighty-one percent of the constructed tax unit–1040 pairings have consistent demographic information based on the CPS and on the 1040 (see **Appendix Table 2**). Taxpayers filing as heads of households, however, were almost as likely to appear as married or unmarried without dependents in the CPS as they were to appear to be unmarried with dependents. In particular, divorced or separated individuals were more likely to claim head-of-household filing status than were individuals who never married, even after controlling for the presence of children in the tax unit. When restricted to constructed tax units in which all members match to the same 1040 and the number of members in the constructed tax unit matches the number of exemptions on the 1040, the filing status is consistent with the demographic characteristics in the CPS in almost all cases.

I further examine the discrepancy between constructed tax units and filing units on the 1040 by comparing the number of exemptions on the 1040 to the number of individuals in the constructed tax unit, among constructed tax units that can be linked to a 1040. The 100.6 million constructed tax units in which every member is linked to the same 1040 consist of 197.6 million individuals, and the 1040s they appear on have a total of 215.3 million exemptions (see **Appendix Table 3**). The excess exemptions appear on 1040s with joint or head-of-household filing status. About 98 percent of individuals on those joint returns are found in constructed tax units that are headed by a married couple, yet even among constructed tax units whose filing status is the same on the 1040 and based on the CPS, there is an excess of 8 million exemptions. Among returns with the head-of-household status, the excess exemptions are found mostly among constructed tax units that consist of unmarried individuals with no dependents in the CPS. (In some cases, that occurs because single taxpayers can claim head-of-household status if they support a parent who lives

apart from the taxpayer, and the parent meets the other criteria for dependents.) The vast majority of CPS individuals are in constructed tax units in which every member can be linked to the same 1040. The remaining individuals appear in constructed tax units in which some but not all members of the tax unit can be linked to a 1040 or are linked to multiple returns. About 68 percent of the individuals in constructed tax units that link to multiple 1040s are in married tax units based on CPS demographic data.

Reported Wages and Salaries

The distribution of wages and salaries based on what is reported on the CPS and on the 1040 is similar (see **Appendix Figure 1**). However, wages and salaries for the same constructed tax unit in the two data sets can be fairly different. Among constructed tax units that have a 1040, about 40 percent of tax units report wage and salary income that differs by 20 percent or more across the two data sets—typically, wage and salary income is higher on the 1040 than on the CPS. The reasons for that discrepancy are not immediately clear.

Tables and Figures
Accompanying This Analysis

Table 1.
Characteristics of Tax Returns for Tax Year 2006, by Presence of a Protected Identification Key for Primary or Secondary Filer

Characteristics	No PIK	Has PIK	All
Average Income (Dollars)			
Wages and salaries	18,326	39,945	39,676
Taxable dividends	61	1,450	1,433
Gross rents and royalties	337	2,171	2,148
Taxable Social Security	48	2,515	2,484
Total income	20,132	57,461	56,996
Adjusted gross income	21,643	58,202	57,746
Presence of 1040 Schedule (Percent)			
Itemized deductions (Schedule A)	16	37	36
Profit or loss from business (Schedule C)	16	16	16
Capital gains and losses (Schedule D)	8	24	24
Supplemental income and loss (Schedule E)	13	13	13
Profit or loss from farming (Schedule F)	0	1	1
Self-employment tax (Schedule SE)	14	12	12
Filing Status (Percent)			
Single	40	45	45
Married filing jointly	35	38	38
Married filing separately	9	2	2
Head of household	16	15	15
Average Number of Dependent Exemptions	1.34	0.62	0.63
Number of Tax Returns (Millions)	1.7	132.6	134.3

Source: Author's calculations, using tax returns provided to the Census Bureau for tax year 2006.

Under the Internal Revenue Code, section 6103(j)(1), the Internal Revenue Service can provide the Census Bureau tax return data for certain purposes. Treasury regulations, however, limit the amount of information to certain items on the tax return. Those data are derived from the Internal Revenue Service's Individual Master File and include all individual income tax returns, including amended returns, for tax year 2006, filed as of October 2007. A Protected Identification Key (PIK) is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources. Tax returns without a PIK are those for which a PIK cannot be assigned to the primary or secondary filer. Constructed tax units are weighted by household weight in the Current Population Survey.

Table 2.
Characteristics of Nondependent Tax Units in the 2007 Current Population Survey, by Presence of a Protected Identification Key

Characteristics	No PIK	Has PIK	All
Average Income (Dollars)			
Wages and salaries	29,503	38,145	37,353
Dividends	474	804	774
Rental income	193	477	451
Social Security	2,205	3,109	3,026
Adjusted gross income	35,976	48,241	47,118
Filing Status Based on CPS Characteristics (Percent)			
Head of household	7	13	12
Married	31	46	45
Single	61	41	43
Transfer Program Receipt (Percent)			
Supplemental Nutrition Assistance Program	5	6	6
Medicare	12	13	13
Medicaid	16	22	21
Average Number of Dependents	0.35	0.67	0.64
Number of Constructed Tax Units (Millions)	13.0	128.7	141.7

Source: Author's calculations, using data from the Census Bureau's 2007 Current Population Survey (CPS).

A Protected Identification Key (PIK) is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources. Publicly available versions of the CPS do not include PIKs. Tax units without a PIK are those for which a PIK cannot be assigned to the primary or secondary filer. Household weights are applied to the primary taxpayer of the constructed tax unit.

Table 3.

Probability of a Constructed Tax Unit Having a Protected Identification Key in Tax Year 2006, by Household's Characteristics and Income Sources

Covariate	Coefficient	Standard Error
Characteristics of Tax Unit		
Composition		
Married, no children	-0.199	0.001
Unmarried, no children	-0.089	0.001
Unmarried, with children	0.342	0.001
Size		
Number under age 18	0.108	0.0004
Number over age 18	0.368	0.0005
Transfer Program Receipt		
Medicaid	0.041	0.001
Medicare	0.064	0.001
Temporary Assistance for Needy Families	0.083	0.002
Supplemental Nutrition Assistance Program	0.124	0.001
Housing assistance	0.111	0.001
Income as Percentage of Federal Poverty Level	0.006	0.00004
Characteristics of Primary Taxpayer		
Age		
30 to 39	-0.040	0.0005
50 to 59	0.138	0.0005
60 to 69	0.290	0.001
70 or older	0.370	0.001
Educational Attainment		
Less than high school	-0.087	0.0005
College	0.067	0.0004
Demographics		
Hispanic	-0.212	0.001
Black	-0.201	0.0005
Other	-0.104	0.001
Citizen	0.549	0.001
Native English speaker	0.275	0.001
Disabled	0.090	0.001
Employment		
Unemployed	-0.191	0.001
In labor force	0.089	0.001
Moved in Last Year	-0.220	0.0004
Respondent Completed March CPS Interview	0.781	0.0004
Constant	-0.593	0.002

Source: Author's calculations, using data from the Census Bureau's 2007 Current Population Survey (CPS).

The dependent variable equals one if any member of the constructed tax unit has a PIK and zero otherwise. Tax units were constructed using the 2007 March CPS. A PIK is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources. Constructed tax units are weighted by CPS household weight. The excluded group consists of constructed tax units with the following characteristics: married with children, and the primary taxpayer is under age 30 or between ages 40 and 49, a high school graduate, and non-Hispanic white.

Table 4.
Number of Observations in Linked Data Before and After Removal of Constructed Tax Units Without a Protected Identification Key and Reweighting
(Number)

	Constructed Tax Units			Individuals in Constructed Tax Unit		
Type of 1040 Match	Unweighted	CPS Household Weight	Reweighted	Unweighted	CPS Household Weight	Reweighted
Without PIK	7,925	12,975,377	n.a.	13,325	20,890,323	n.a.
With PIK						
No member is linked to 1040	13,277	20,832,417	23,722,434	24,682	36,611,575	40,662,564
At least one member is linked to 1040	71,345	107,857,251	117,930,746	168,632	235,209,228	251,949,903
Total	92,547	141,665,046	141,653,180	206,639	292,711,126	292,612,467

Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns for tax year 2006.

A PIK is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources.

n.a. = not applicable.

Table 5.
Constructed Tax Units and Individuals in Tax Year 2006 in the Linked Data, by Filing and Current Population Survey Demographic Characteristics

	Tax Units With 1040			Tax Units Without 1040		
Characteristics	Number (Millions)	Share (Percent)	Number of Individuals (Millions)	Number (Millions)	Share (Percent)	Number of Individuals (Millions)
Age 65 or Older	17.7	15	28.0	7.4	31	10.0
Unmarried With no Dependents	39.0	33	39.0	8.8	37	8.8
Unmarried With Dependents	12.6	11	33.0	3.5	15	9.4
Married With no Dependents	19.7	17	38.2	1.6	7	3.0
Married With Dependents	29.0	25	113.7	2.4	10	9.5
Total	117.9	100	251.9	23.7	100	40.7

Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns.

Constructed tax units in which no member has a Protected Identification Key (PIK) are excluded. A PIK is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources.

Table 6.
Sources of Income and Average Amounts for Constructed Tax Units With 1040 in Tax Year 2006

Share of Group With Income Source
(Percent)

Income	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
Taxable						
Adjusted gross income	88	97	94	98	99	96
Wages and salaries	27	91	88	90	96	82
Self-employment income	5	6	5	13	13	9
Dividends	30	15	9	32	28	22
Interest	64	39	27	62	56	49
Rental income	11	4	2	10	7	6
Unemployment insurance	1	3	5	5	5	4
Partially Taxable						
Social Security	91	3	5	14	3	18
Nontaxable						
Supplemental Security Income	1	1	2	1	1	1
Public assistance	0	0	4	0	1	1
Housing assistance	2	2	9	0	1	2
Supplemental Nutrition Assistance Program	1	3	18	1	4	4
Medicare	96	2	7	9	4	18
Medicaid	7	4	38	4	17	11

Average Amount[a]
(Dollars)

Income	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	**Memorandum:** Total Amount (Billions)
Taxable							
Adjusted gross income	32,000	37,000	31,200	84,800	90,800	56,900	6,705
Wages and salaries	10,200	33,100	28,500	70,100	81,900	47,400	5,587
Self-employment income	1,700	1,600	1,300	5,000	5,100	3,000	355
Dividends	2,100	400	200	1,600	800	900	109
Interest	4,000	900	500	2,900	1,700	1,800	218
Rental income	1,100	300	100	800	600	600	66
Unemployment insurance	30	100	200	200	200	200	18
Partially Taxable							
Social Security	15,000	400	500	2,000	400	2,900	337
Nontaxable							
Supplemental Security Income	100	100	100	100	100	100	9
Public assistance	<5	<5	100	<5	20	20	3
Supplemental Nutrition Assistance Program	20	50	500	10	100	100	12

Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns.

Constructed tax units in which no member has a Protected Identification Key (PIK) are excluded. A PIK is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources. Income and demographic characteristics are based on information reported in the CPS.

a. Average is across all units, including those with zero income in a category.

Table 7.
Sources of Income and Average Amounts for Constructed Tax Units Without 1040 in Tax Year 2006

Share of Group With Income Source
(Percent)

Income	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
Taxable						
Adjusted gross income	51	67	59	74	89	64
Wages and salaries	5	52	47	58	80	40
Self-employment income	1	8	6	12	15	6
Dividends	7	4	3	13	13	6
Interest	32	17	11	28	32	23
Rental income	2	2	1	5	6	3
Unemployment insurance	0	3	3	3	3	2
Partially Taxable						
Social Security	90	16	12	25	7	38
Nontaxable						
Supplemental Security Income	9	12	12	10	5	10
Public assistance	0	1	13	1	2	3
Housing assistance	13	11	18	4	3	11
Supplemental Nutrition Assistance Program	10	16	37	11	12	17
Medicare	97	14	15	23	8	40
Medicaid	24	22	59	18	36	29

Average Amount[a]
(Dollars)

Income	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	**Memorandum:** Total Amount (Billions)
Taxable							
Adjusted gross income	6,200	17,200	13,400	44,400	60,900	19,500	463
Wages and salaries	1,300	13,600	10,900	33,600	51,600	14,600	347
Self-employment income	400	1,900	1,500	5,700	5,900	2,000	48
Dividends	200	100	100	800	600	200	6
Interest	1,100	300	100	1,300	1,400	700	17
Rental income	200	100	50	500	600	200	5
Unemployment insurance	<5	100	100	200	100	100	2
Partially Taxable							
Social Security	11,500	1,600	1,200	3,500	1,100	4,700	112
Nontaxable							
Supplemental Security Income	500	800	800	900	300	600	15
Public assistance	10	30	500	30	100	100	2
Supplemental Nutrition Assistance Program	100	200	1,100	200	400	300	8

Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns.

Constructed tax units in which no member has a Protected Identification Key (PIK) are excluded. A PIK is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources. Income and demographic characteristics are based on information reported in the CPS.

a. Average is across all units, including those with zero income in a category.

Table 8.
Adjustments to Public Use File and Linked Data for Tax Year 2006

Number of Filers After Adjustments to Public Use File	Public Use File, All Nondependent Returns		Public Use File, After Removing Foreign and Late Returns	
	Number (Millions)	Share (Percent)	Number (Millions)	Share (Percent)
Age 65 or Older	18.8	15	15.4	13
Unmarried With No Dependents	43.5	34	40.8	35
Unmarried With Dependents	22.5	18	19.3	16
Married With No Dependents	15.7	12	16.4	14
Married With Dependents	27.5	21	25.6	22
Total	128.0	100	117.5	100

Number of Nonfilers After Adjustments to Linked Data	Linked Data		Linked Data Without Late Filers	
	Number (Millions)	Share (Percent)	Number (Millions)	Share (Percent)
Age 65 or Older	7.4	31	7.0	34
Unmarried With No Dependents	8.8	37	8.1	39
Unmarried With Dependents	3.5	15	2.9	14
Married With No Dependents	1.6	7	1.2	6
Married With Dependents	2.4	10	1.6	8
Total	23.7	100	20.8	100

Source: Author's calculations, using data from the Internal Revenue Service's Public Use File (PUF) and the Census Bureau's linked Current Population Survey and tax returns.

The number of tax returns in the PUF excludes returns filed by dependents. Foreign returns in the PUF are those filed by taxpayers who may be abroad (based on having an address outside of the 50 states, claiming the foreign tax credit, or owing foreign earned income tax). Late returns were filed in 2007 for the 2003–2005 tax years. The 2006 PUF does not contain age information. Age was imputed on the basis of the presence of Social Security and pension income.

Table 9.
Sources of Income and Average Amounts for Constructed Tax Units Without 1040, After Removing Potential Late Filers, in Tax Year 2006

Share of Group With Income Source
(Percent)

Income	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
Taxable						
Adjusted gross income	51	65	56	70	88	61
Wages and salaries	4	49	43	53	79	36
Self-employment income	1	8	6	13	15	6
Dividends	6	4	2	12	12	6
Interest	32	16	11	26	28	22
Rental income	3	2	1	6	5	2
Unemployment insurance	0	3	2	3	3	2
Partially Taxable						
Social Security	90	17	13	27	8	41
Nontaxable						
Supplemental Security Income	9	13	13	11	5	11
Public assistance	0	1	14	1	2	3
Housing assistance	13	11	19	5	3	12
Supplemental Nutrition Assistance Program	10	17	40	11	12	17
Medicare	97	15	15	24	9	43
Medicaid	24	24	62	19	38	30

Average Amount[a]
(Dollars)

Income	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	**Memorandum:** Total Amount (Billions)
Taxable							
Adjusted gross income	5,600	15,300	11,400	41,400	56,100	16,100	336
Wages and salaries	900	11,600	8,800	29,900	46,600	11,400	236
Self-employment income	300	1,900	1,600	6,200	6,500	2,000	41
Dividends	200	100	100	900	400	200	4
Interest	1,100	300	100	1,500	1,100	700	14
Rental income	200	100	100	500	500	200	4
Unemployment insurance	<5	100	100	100	200	100	2
Partially Taxable							
Social Security	11,600	1,700	1,300	3,700	1,200	5,100	105
Nontaxable							
Supplemental Security Income	500	800	900	900	400	700	14
Public assistance	10	30	500	30	100	100	2
Supplemental Nutrition Assistance Program	100	200	1,200	200	400	300	7

Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns.

Constructed tax units in which no member has a Protected Identification Key (PIK) are excluded. A PIK is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources. Late filers are imputed on the basis of wage income and filing status. Income and demographic characteristics are based on information reported in the CPS.

a. Average is across all units, including those with zero income in a category.

Table 10.
Simulated Filers and Nonfilers in Tax Year 2006 Compared With Filers and Nonfilers in Linked Data, by Current Population Survey Demographic Characteristics

	Filer in Linked Data		Simulated Filer Using Predicted Income		Simulated Filer Using Predicted Probability	
Simulated Filers	Number (Millions)	Share (Percent)	Number (Millions)	Share (Percent)	Number (Millions)	Share (Percent)
Age 65 or Older	17.7	15	17.7	14	17.8	15
Unmarried With No Dependents	39.0	33	46.7	36	38.8	33
Unmarried With Dependents	12.6	11	15.8	12	12.4	11
Married With No Dependents	19.7	17	18.6	15	19.6	17
Married With Dependents	29.0	25	29.3	23	28.9	25
Total	117.9	100	128.1	100	117.5	100

	Nonfiler in Linked Data		Simulated Nonfiler Using Predicted Income		Simulated Nonfiler Using Predicted Probability	
Simulated Nonfilers	Number (Millions)	Share (Percent)	Number (Millions)	Share (Percent)	Number (Millions)	Share (Percent)
Age 65 or Older	7.4	31	7.3	54	7.2	30
Unmarried With No Dependents	8.8	37	1.2	9	9.1	38
Unmarried With Dependents	3.5	15	0.2	1	3.6	15
Married With No Dependents	1.6	7	2.7	20	1.7	7
Married With Dependents	2.4	10	2.2	16	2.6	11
Total	23.7	100	13.6	100	24.1	100

Source: Author's calculations, using data from the Internal Revenue Service's Public Use File, the Census Bureau's Current Population Survey (CPS), and the Census Bureau's linked CPS and tax returns.

Constructed tax units in which no member has a Protected Identification Key (PIK) are excluded from the linked data. A PIK is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources. All constructed tax units in the publicly available CPS, which does not include PIKs, are included in the simulations.

Table 11.
Sources of Income and Average Amounts for Simulated Filers Using Predicted Income in Tax Year 2006

Share of Group With Income Source
(Percent)

	Simulated Filers Using Predicted Income						Difference Relative to Filers in Linked Data (Table 11 Compared With Table 6)					
Income	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
Taxable												
Adjusted gross income	98	94	87	99	100	96	10	-3	-7	1	1	0
Wages and salaries	30	85	79	92	96	80	3	-6	-9	2	0	-2
Self-employment income	6	6	5	13	14	9	1	0	0	0	1	0
Dividends	32	13	7	33	28	21	2	-2	-2	1	0	-1
Interest	69	35	24	63	56	47	5	-4	-3	1	0	-2
Rental income	11	3	2	10	7	6	0	-1	0	0	0	0
Unemployment insurance	1	3	4	5	5	4	0	0	-1	0	0	0
Partially Taxable												
Social Security	90	5	6	11	2	17	-1	2	1	-3	-1	-1
Nontaxable												
Supplemental Security Income	1	2	4	1	1	2	0	1	2	0	0	1
Public assistance	0	0	6	0	0	1	0	0	2	0	-1	0
Housing assistance	2	3	11	1	1	3	0	1	2	1	0	1
Supplemental Nutrition Assistance Program	1	5	23	1	3	5	0	2	5	0	-1	1
Medicare	95	4	9	8	4	17	-1	2	2	-1	0	-1
Medicaid	8	6	42	3	16	13	1	2	4	-1	-1	2

Average Amount[a]
(Dollars)

	Simulated Filers Using Predicted Income							Difference Relative to Filers in Linked Data (Table 11 Compared With Table 6)						
Income	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	**Memorandum:** Total Amount (Billions)	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	Difference in Total Income (Billions)
Taxable														
Adjusted gross income	35,200	34,000	27,600	91,100	94,500	55,500	7,105	3,200	-3,000	-3,600	6,300	3,600	-1,400	400
Wages and salaries	11,400	30,000	24,900	75,200	84,800	45,900	5,875	1,200	-3,100	-3,600	5,000	2,900	-1,500	288
Self-employment income	1,900	1,700	1,300	5,600	5,600	3,200	405	200	100	50	600	500	200	50
Dividends	2,200	300	200	1,800	900	900	115	100	-20	-20	200	40	-30	6
Interest	4,500	800	400	3,200	1,800	1,800	235	500	-90	-60	300	100	-10	17
Rental income	1,200	200	100	900	600	500	69	40	-30	-20	50	40	-20	3
Unemployment insurance	30	100	200	200	200	100	19	*	-10	-10	-10	-10	-10	1
Partially Taxable														
Social Security	14,800	500	700	1,600	300	2,600	337	-200	200	100	-400	-100	-200	0
Nontaxable														
Supplemental Security Income	100	100	300	100	100	100	15	*	100	100	-40	-20	40	6
Public assistance	10	10	200	*	20	30	4	*	*	100	*	*	10	1
Supplemental Nutrition Assistance Program	20	100	600	10	100	100	16	*	30	100	*	-20	30	4

Source: Author's calculations, using data from the Internal Revenue Service's Public Use File and the Census Bureau's Current Population Survey (CPS).

Income and demographic characteristics are based on information reported in the CPS. * = between -$5 and $5.

a. Average is across all units, including those with zero income in a category.

Table 12.
Sources of Income and Average Amounts for Simulated Nonfilers Using Predicted Income in Tax Year 2006

Share of Group With Income Source
(Percent)

Income	Simulated Nonfilers Using Predicted Income: Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	Difference Relative to Nonfilers in Linked Data (Table 12 Compared With Table 7): Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
Taxable												
Adjusted gross income	26	0	1	74	79	42	-25	-67	-58	0	-10	-22
Wages and salaries	1	0	2	52	67	22	-4	-52	-45	-6	-13	-18
Self-employment income	0	0	4	10	10	4	-1	-8	-2	-2	-5	-2
Dividends	2	0	0	7	3	3	-5	-4	-3	-6	-10	-3
Interest	20	0	0	23	13	18	-12	-17	-11	-5	-19	-5
Rental income	1	0	1	2	2	1	-1	-2	0	-3	-4	-2
Unemployment insurance	0	0	0	5	5	2	0	-3	-3	2	2	0
Partially Taxable												
Social Security	91	30	18	39	17	62	82	18	6	29	12	52
Nontaxable												
Supplemental Security Income	9	21	15	9	8	10	-4	10	-3	5	5	-1
Public assistance	0	3	22	1	5	2	-10	-13	-15	-10	-7	-15
Housing assistance	13	17	13	3	7	10	-84	3	-2	-20	-1	-30
Supplemental Nutrition Assistance Program	10	22	32	8	23	13	-14	0	-27	-10	-13	-16
Medicare	98	26	15	29	17	64	1	12	0	6	9	24
Medicaid	22	43	65	20	53	29	-2	21	6	2	17	0

Average Amount[a]
(Dollars)

Income	Simulated Nonfilers Using Predicted Income: Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	**Memorandum:** Total Amount (Billions)	Difference Relative to Nonfilers in Linked Data (Table 12 Compared With Table 7): Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	Difference in Total Income (Billions)
Taxable														
Adjusted gross income	200	60	30	11,300	11,800	4,300	58	-6,000	-17,300	-13,500	-32,900	-47,600	-15,300	-405
Wages and salaries	10	0	20	8,400	10,000	3,300	45	-1,300	-13,700	-10,800	-25,200	-40,500	-11,400	-302
Self-employment income	-20	0	-200	1,000	800	300	4	-400	-1,900	-1,700	-4,700	-4,800	-1,700	-44
Dividends	0	0	0	50	20	20	0	-200	-100	-100	-800	-500	-200	-6
Interest	50	0	*	200	100	100	1	-1,100	-300	-100	-1,200	-1,300	-700	-16
Rental income	0	0	-100	*	20	*	0	-200	-100	-100	-500	-100	-200	-5
Unemployment insurance	*	0	0	200	200	100	1	0	-100	-100	20	-800	-10	-1
Partially Taxable														
Social Security	11,700	3,100	2,700	6,100	2,500	8,200	112	200	1,500	1,400	2,100	2,100	3,500	0
Nontaxable														
Supplemental Security Income	500	1,400	1,200	800	600	600	9	0	600	400	-100	100	0	-6
Public assistance	10	70	1,100	40	200	100	1	0	40	600	10	-200	-40	-1
Supplemental Nutrition Assistance Program	100	300	1,000	100	700	200	3	0	100	-100	-40	600	-100	-5

Source: Author's calculations, using data from the Internal Revenue Service's Public Use File and the Census Bureau's Current Population Survey (CPS).

Income and demographic characteristics are based on information reported in the CPS. * = between -$5 and $5.

a. Average is across all units, including those with zero income in a category.

Table 13.
Average Income Reported by Filers in the Linked Data for Tax Year 2006,
by Filing Status Imputed Using Predicted Income Method and Demographic Characteristics

Simulated Filers With 1040 in Linked Data	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
			Income Reported on 1040 (Dollars)			
Adjusted Gross Income	76,800	37,000	30,200	93,600	98,100	65,600
Wages and Salaries	12,000	30,400	26,100	70,600	79,600	46,100
Taxable Dividends	4,600	500	200	1,600	1,600	1,400
Gross Rents and Royalties	4,000	1,000	400	3,200	2,600	2,100
			Income Reported in CPS (Dollars)			
Adjusted Gross Income	38,700	37,300	31,300	93,000	95,200	60,000
Wages and Salaries	12,400	33,400	28,600	77,100	86,000	50,000
Dividends	2,500	400	200	1,800	900	1,000
Rental Income	1,400	300	100	900	600	600
Memorandum:						
Number of Tax Units (Millions)	14.6	38.7	12.6	17.7	27.3	110.8
Share of Simulated Filers (Percent)	82	83	80	95	93	87

Simulated Nonfilers With 1040 in Linked Data	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
			Income Reported on 1040 (Dollars)			
Adjusted Gross Income	34,800	23,900	14,800	38,500	45,500	37,800
Wages and Salaries	1,800	17,500	10,500	18,900	19,800	11,500
Taxable Dividends	2,800	100	100	700	2,400	2,000
Gross Rents and Royalties	1,800	600	1,800	1,600	1,100	1,600
			Income Reported in CPS (Dollars)			
Adjusted Gross Income	300	10	0	13,700	12,800	7,000
Wages and Salaries	10	0	50	10,300	10,900	5,500
Dividends	10	0	0	100	30	30
Rental Income	10	0	-200	-10	30	*
Memorandum:						
Number of Tax Units (Millions)	3.1	0.3	0.1	2.0	1.6	7.0
Share of Simulated Nonfilers (Percent)	42	22	34	74	75	52

Source: Author's calculations, using data from the Internal Revenue Service's Public Use File, the Census Bureau's Current Population Survey (CPS), and the Census Bureau's linked CPS and tax returns.

Average is across all units, including those with zero income in a category. Demographic characteristics are based on information reported in the CPS. * = between -$5 and $5.

Table 14.
Estimates of the Probability of Filing Form 1040 in Tax Year 2006, by Characteristics of Primary Taxpayer

A. Unmarried, Under Age 65, Nondependent Filers

	No Dependents		1 Dependent		2 or More Dependents	
Variable	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
Log (Gross income)	0.125***	0.010	0.15***	0.020	0.182***	0.021
Negative Gross Income Indicator	0.452***	0.096	0.599***	0.179	1.045***	0.184
Income Sources						
Wages and salaries	0.669***	0.046	0.603***	0.101	0.715***	0.092
Interest	0.211***	0.027	0.243***	0.059	0.143**	0.063
Dividends	0.14***	0.043	-0.025	0.097	-0.028	0.103
Self-employment income	-0.108**	0.049	-0.13	0.111	-0.018	0.104
Rental income	-0.067	0.066	-0.2	0.146	0.15	0.169
Retirement income	0.156**	0.072	0.144	0.189	0.162	0.232
Social Security	-0.114**	0.045	-0.091	0.085	0.026	0.078
Means-Tested Transfers	-0.276***	0.036	0.065	0.057	-0.047	0.055
Number With Medicaid Coverage	-0.273***	0.038	-0.062**	0.031	-0.015	0.016
Demographic Characteristics						
Less than high school education	-0.263***	0.032	-0.411***	0.054	-0.21***	0.053
College education	0.201***	0.028	0.03	0.067	0.09	0.073
Black	-0.283***	0.028	-0.111**	0.053	-0.033	0.053
Hispanic	-0.095***	0.030	-0.145***	0.055	-0.16***	0.055
Other	-0.095***	0.037	-0.162**	0.077	-0.06	0.080
Intercept	-0.696***	0.096	-0.869***	0.186	-1.165***	0.194
Number of Observations	24,634		5,594		5,931	
Mean of Dependent Variable (Weighted)	**0.82**		**0.77**		**0.79**	

B. Unmarried, Age 65 or Older, Nondependent Filers

	No Dependents		1 Dependent		2 or More Dependents	
Variable	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
Log (Gross income)	0.135***	0.013	0.171***	0.052	0.197***	0.067
Negative Gross Income Indicator	0.63***	0.108	1.113***	0.428	1.43**	0.592
Income Sources						
Wages and salaries	0.552***	0.067	0.871***	0.304	0.18	0.348
Interest	0.179***	0.049	0.385*	0.197	0.47	0.306
Dividends	0.329***	0.058	0.268	0.288	-0.771**	0.353
Self-employment income	0.028	0.131	-1.218*	0.688	0.403	0.644
Rental income	0.467***	0.088	0.708**	0.333	-0.665	0.428
Retirement income	0.074	0.050	0.02	0.221	-0.104	0.287
Social Security	0.118*	0.065	-0.11	0.181	-0.038	0.246
Means-Tested Transfers	-0.44***	0.052	-0.307*	0.175	-0.229	0.227
Number With Medicaid Coverage	-0.368***	0.057	-0.229**	0.109	-0.082	0.086
Demographic Characteristics						
Less than high school education	-0.358***	0.041	-0.472***	0.144	-0.266	0.210
College education	0.197***	0.056	0.378	0.255	-0.033	0.306
Black	-0.26***	0.052	-0.135	0.158	-0.284	0.220
Hispanic	-0.148**	0.071	-0.458**	0.210	-0.776**	0.342
Other	-0.125*	0.076	-0.26	0.259	-0.019	0.322
Intercept	-0.825***	0.127	-0.834*	0.465	-0.799	0.636
Number of Observations	6,616		496		254	
Mean of Dependent Variable (Weighted)	**0.62**		**0.60**		**0.63**	

C. Married, Under Age 65, Nondependent Filers

	No Dependents		1 Dependent		2 or More Dependents	
Variable	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
Log (Gross income)	0.149***	0.021	0.101***	0.030	0.111***	0.020
Negative Gross Income Indicator	0.62***	0.197	0.061	0.302	0.526**	0.214
Income Sources						
Wages and salaries	0.405***	0.074	0.321***	0.114	0.467***	0.074
Interest	0.301***	0.048	0.201***	0.058	0.166***	0.038
Dividends	0.023	0.057	0.053	0.072	0.042	0.046
Self-employment income	-0.108*	0.060	-0.238***	0.073	-0.07	0.048
Rental income	-0.084	0.076	-0.097	0.103	-0.189***	0.060
Retirement income	0.2***	0.073	0.113	0.158	0.1	0.124
Social Security	0.054	0.059	-0.235**	0.097	0.01	0.089
Means-Tested Transfers	-0.509***	0.088	-0.154	0.105	-0.062	0.061
Number With Medicaid Coverage	-0.105**	0.042	-0.05	0.033	-0.02*	0.012
Demographic Characteristics						
Less than high school education	-0.17**	0.073	-0.33***	0.082	-0.371***	0.051
College education	-0.01	0.046	-0.017	0.057	0.027	0.039
Black	-0.289***	0.061	-0.21**	0.083	-0.119**	0.061
Hispanic	-0.1*	0.056	-0.417***	0.058	-0.518***	0.038
Other	-0.072	0.062	-0.118	0.075	-0.038	0.056
Intercept	-0.44**	0.200	0.282	0.301	0.005	0.210
Number of Observations	11,081		7,182		17,041	
Mean of Dependent Variable (Weighted)	**0.92**		**0.92**		**0.92**	

D. Married, Age 65 or Older, Nondependent Filers

	No Dependents		1 Dependent		2 or More Dependents	
Variable	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
Log (Gross income)	0.14***	0.017	0.089	0.076	0.15	0.100
Negative Gross Income Indicator	0.72***	0.148	-0.491	0.618	0.15	0.871
Income Sources						
Wages and salaries	0.492***	0.077	0.499	0.321	0.489	0.353
Interest	0.158**	0.062	-0.327	0.241	0.101	0.330
Dividends	0.452***	0.074	0.842**	0.364	-0.13	0.393
Self-employment income	-0.006	0.121	0.383	0.532	-0.495	0.403
Rental income	0.21**	0.100	-0.004	0.390	0.023	0.514
Retirement income	-0.193***	0.068	-0.276	0.312	-0.67*	0.348
Social Security	-0.008	0.090	0.391	0.268	0.232	0.335
Means-Tested Transfers	-0.648***	0.102	-0.027	0.276	-0.364	0.361
Number With Medicaid Coverage	-0.216***	0.045	-0.292***	0.105	0.064	0.107
Demographic Characteristics						
Less than high school education	-0.605***	0.064	-0.727***	0.227	-0.121	0.342
College education	0.308***	0.064	-0.073	0.237	0.567	0.404
Black	-0.204**	0.080	-0.294	0.237	-0.532*	0.320
Hispanic	-0.21***	0.079	-0.361	0.253	-0.436	0.388
Other	-0.291***	0.083	-0.111	0.265	-0.561	0.372
Intercept	-0.26	0.172	0.485	0.671	-0.182	0.990
Number of Observations	5,237		362		194	
Mean of Dependent Variable (Weighted)	**0.83**		**0.77**		**0.80**	

E. Dependent Filers

	Age 65 or Older		Under Age 65	
Variable	Coefficient	Standard Error	Coefficient	Standard Error
Log (Gross income)	0.056	0.054	0.107***	0.011
Negative Gross Income Indicator	0.372	0.430	0.108	0.080
Income Sources				
Wages and salaries	0.75	0.467	0.9***	0.046
Interest	0.183	0.234	0.12***	0.037
Dividends	0.84***	0.280	0.274***	0.062
Self-employment income	.	.	-0.026	0.109
Rental income	0.315	0.411	0.672***	0.255
Social Security	-0.052	0.124	-0.044	0.058
Means-Tested Transfers	-0.35**	0.168	-0.074**	0.033
Number With Medicaid Coverage	-0.201	0.132	-0.356***	0.027
Demographic Characteristics				
Less than high school education	-0.321***	0.104	-1.171***	0.023
College education	0.119	0.175	0.423***	0.079
Black	-0.297**	0.133	-0.142***	0.030
Hispanic	-0.635***	0.156	-0.127***	0.025
Other	-0.234	0.146	-0.11***	0.031
Intercept	-0.739*	0.449	-0.67***	0.083
Number of Observations	909		62,438	
Mean of Dependent Variable (Weighted)	**0.23**		**0.10**	

Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns.

The dependent variable equals one if the primary taxpayer has a 1040. Constructed tax units in which no member has a Protected Identification Key (PIK) are excluded. A PIK is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources. Income and demographic characteristics are based on information reported in the CPS. Gross income is calculated as total income net of taxable Social Security income. Income source variables are indicators denoting whether a source was present. Means-tested transfers include Temporary Assistance for Needy Families, Supplemental Nutrition Assistance Program, Low Income Home Energy Assistance Program, and housing assistance. The excluded group consists of constructed tax units with the following characteristics: the primary taxpayer is non-Hispanic white with more than a college education.

*** p<0.01, ** p<0.05, * p<0.1.

Table 15.
Sources of Income and Average Amounts for Simulated Filers Using Predicted Probability in Tax Year 2006

Share of Group with Income Source
(Percent)

Income	Simulated Filers Using Predicted Probability: Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	Difference Relative to Filers in Linked Data (Table 15 Compared With Table 6): Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
Taxable												
Adjusted gross income	96	100	100	100	100	99	8	3	6	2	1	3
Wages and salaries	30	97	97	93	97	86	3	6	9	3	1	4
Self-employment income	6	4	4	13	13	8	1	-2	-1	0	0	-1
Dividends	32	15	9	32	29	23	2	0	0	0	1	1
Interest	70	39	29	62	58	51	6	0	2	0	2	2
Rental income	11	4	2	10	7	7	0	0	0	0	0	1
Unemployment insurance	1	3	5	5	5	4	0	0	0	0	0	0
Partially Taxable												
Social Security	90	2	3	13	3	17	-1	-1	-2	-1	0	-1
Nontaxable												
Supplemental Security Income	0	0	1	1	1	1	-1	-1	-1	0	0	0
Public assistance	0	0	3	0	0	0	0	0	-1	0	-1	-1
Housing assistance	1	1	8	0	1	1	-1	-1	-1	0	0	-1
Supplemental Nutrition Assistance Program	1	1	16	0	3	3	0	-2	-2	-1	-1	-1
Medicare	95	1	7	8	4	18	-1	-1	0	-1	0	0
Medicaid	5	2	35	3	15	9	-2	-2	-3	-1	-2	-2

Average Amount[a]
(Dollars)

Income	Simulated Filers Using Predicted Probability: Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	**Memorandum:** Total Amount (Billions)	Difference Relative to Filers in Linked Data (Table 15 Compared With Table 6): Age 65 or Oler	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	Difference in Total Income (Billions)
Taxable														
Adjusted gross income	34,600	39,100	34,100	87,400	94,800	59,600	7,011	2,600	2,100	2,900	2,600	3,900	2,800	306
Wages and salaries	11,300	35,500	31,300	72,200	85,300	49,800	5,849	1,100	2,400	2,800	2,100	3,400	2,400	262
Self-employment income	1,900	1,200	1,300	5,200	5,400	3,000	355	200	-400	*	200	300	10	0
Dividends	2,200	400	200	1,700	900	1,000	115	100	30	20	100	100	100	6
Interest	4,400	1,000	500	3,000	1,800	2,000	233	400	40	40	100	100	100	15
Rental income	1,200	300	100	800	600	600	68	30	-10	*	*	40	10	2
Unemployment insurance	30	100	200	200	200	200	18	*	-10	10	-10	-10	-10	0
Partially Taxable														
Social Security	15,000	200	300	1,800	300	2,700	323	-100	-200	-200	-200	-100	-100	-14
Nontaxable														
Supplemental Security Income	30	10	100	100	100	40	5	-30	-40	-60	-40	-20	-40	-4
Public assistance	*	*	100	*	20	10	2	*	*	-40	*	-10	-10	-1
Supplemental Nutrition Assistance Program	10	20	400	*	100	100	9	-10	-30	-60	-10	-20	-20	-3

Source: Author's calculations, using data from the Internal Revenue Service's Public Use File and the Census Bureau's Current Population Survey (CPS).

Income and demographic characteristics are based on information reported in the CPS. * = between -$5 and $5.

a. Average is across all units, including those with zero income in a category.

Table 16.
Sources of Income and Average Amounts for Simulated Nonfilers Using Predicted Probability in Tax Year 2006

Share of Group with Income Source
(Percent)

Income	Simulated Nonfilers Using Predicted Probability: Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	Difference Relative to Nonfilers in Linked Data (Table 16 Compared With Table 7): Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
Taxable												
Adjusted gross income	30	54	35	52	78	46	-21	-13	-24	-22	-11	-18
Wages and salaries	1	21	15	20	60	18	-4	-31	-32	-38	-20	-22
Self-employment income	1	16	9	15	16	10	0	8	3	3	1	4
Dividends	1	2	1	2	1	1	-6	-2	-2	-11	-12	-5
Interest	18	12	6	11	8	12	-14	-5	-5	-17	-24	-11
Rental income	1	2	1	3	3	2	-1	0	0	-2	-3	-1
Unemployment insurance	0	2	2	3	4	2	0	-1	-1	0	1	0
Partially Taxable												
Social Security	90	23	18	39	12	42	0	7	6	14	5	4
Nontaxable												
Supplemental Security Income	10	13	15	14	6	12	1	1	3	4	1	2
Public assistance	0	2	16	2	4	4	0	1	3	1	2	1
Housing assistance	16	15	22	9	6	15	3	4	4	5	3	4
Supplemental Nutrition Assistance Program	12	23	45	18	19	22	2	7	8	7	7	5
Medicare	98	19	18	36	13	43	1	5	3	13	5	3
Medicaid	28	30	70	29	55	38	4	8	11	11	19	9

Average Amount[a]
(Dollars)

Income	Simulated Nonfilers Using Predicted Probability: Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	**Memorandum:** Total Amount (Billions)	Difference Relative to Nonfilers in Linked Data (Table 16 Compared With Table 7): Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All	Difference in Total Income (Billions)
Taxable														
Adjusted gross income	1,100	7,500	3,100	6,600	21,000	6,300	153	-5,100	-9,700	-10,300	-37,800	-39,900	-13,200	-310
Wages and salaries	100	2,300	1,000	2,900	15,900	2,900	71	-1,200	-11,300	-9,900	-30,700	-35,700	-11,700	-276
Self-employment income	100	3,600	1,400	2,700	4,400	2,300	55	-300	1,800	-100	-3,000	-1,400	300	7
Dividends	10	40	10	100	*	20	1	-200	-100	-100	-800	-600	-200	-5
Interest	100	200	40	100	30	100	3	-1,000	-200	-100	-1,300	-1,300	-600	-14
Rental income	10	100	10	40	100	100	2	-100	10	-30	-400	-400	-100	-3
Unemployment insurance	*	100	100	100	200	100	2	*	*	-20	-100	100	*	0
Partially Taxable														
Social Security	11,400	2,500	1,900	6,200	1,800	5,200	126	-200	900	700	2,700	600	500	14
Nontaxable														
Supplemental Security Income	500	900	1,000	1,300	500	800	19	100	100	200	400	200	100	4
Public assistance	10	100	600	100	200	100	3	*	20	100	40	100	40	1
Supplemental Nutrition Assistance Program	100	300	1,300	300	600	400	11	20	100	200	100	200	100	3

Source: Author's calculations, using data from the Internal Revenue Service's Public Use File and the Census Bureau's Current Population Survey (CPS).

Income and demographic characteristics are based on information reported in the CPS. * = between -$5 and $5.

a. Average is across all units, including those with zero income in a category.

Table 17.
Average Income Reported by Filers in the Linked Data for Tax Year 2006,
by Filing Status Imputed Using Predicted Probability Method and Demographic Characteristics

Simulated Filers With 1040 in Linked Data	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
			Income Reported on 1040 (Dollars)			
Adjusted Gross Income	76,900	39,000	31,900	91,200	99,000	68,100
Wages and Salaries	11,700	32,600	28,300	68,000	79,400	47,400
Taxable Dividends	4,600	400	200	1,500	1,800	1,500
Gross Rents and Royalties	4,000	1,000	400	3,100	2,600	2,100
			Income Reported in CPS (Dollars)			
Adjusted Gross Income	37,600	40,100	34,700	88,900	94,700	61,700
Wages and Salaries	12,100	36,700	32,100	73,700	85,600	51,800
Dividends	2,500	400	200	1,700	900	1,000
Rental Income	1,400	300	200	900	600	600
Memorandum:						
Number of Tax Units (Millions)	14.9	34.8	11.1	18.7	27.4	106.8
Share of Simulated Filers (Percent)	84	90	89	95	95	91

Simulated Nonfilers With 1040 in Linked Data	Age 65 or Older	Unmarried With No Dependents	Unmarried With Dependents	Married With No Dependents	Married With Dependents	All
			Income Reported on 1040 (Dollars)			
Adjusted Gross Income	29,100	19,400	17,300	27,100	28,800	23,600
Wages and Salaries	2,000	11,200	10,200	15,200	22,600	10,900
Taxable Dividends	2,300	700	300	400	100	900
Gross Rents and Royalties	1,300	900	500	1,400	1,400	1,100
			Income Reported in CPS (Dollars)			
Adjusted Gross Income	1,600	11,300	4,200	9,200	22,600	9,500
Wages and Salaries	200	4,100	1,400	4,400	17,100	4,800
Dividends	10	40	10	40	*	20
Rental Income	10	300	10	100	100	200
Memorandum:						
Number of Tax Units (Millions)	2.7	4.2	1.5	1.0	1.7	11.1
Share of Simulated Nonfilers (Percent)	38	46	42	57	65	46

Source: Author's calculations, using data from the Internal Revenue Service's Public Use File, the Census Bureau's Current Population Survey (CPS), and the Census Bureau's linked CPS and tax returns.

Average is across all units, including those with zero income in a category. Demographic characteristics are based on information reported in the CPS. * = between -$5 and $5.

Appendix Table 1.
Constructed Tax Units, by Number of Matched 1040s and Relative Size of Tax Units in Tax Year 2006

(Number, in millions)

	Size of Filing Unit on 1040 Relative to Size of Constructed Tax Unit			
Number of Matched 1040s	Same	Larger Than	Smaller Than	Total
No Matching 1040	n.a.	n.a.	n.a.	23.7
Only One 1040 Matches				
All members linked to the same 1040	84.4	14.4	1.8	100.6
Some members do not appear on 1040	1.9	0.5	7.0	9.3
More Than One 1040 Matches				
All members linked to 1040s	2.8	4.4	0.1	7.2
Some members do not appear on 1040s	0.2	0.2	0.4	0.8
Total	89.3	19.5	9.2	141.7

Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns.

The sample excludes constructed tax units in which no member has a Protected Identification Key (PIK). A PIK is a unique identifier assigned by the Census Bureau using a probabilistic match to enable linkages across data sources. The size of the filing unit on the 1040 is determined by the number of personal and dependent exemptions claimed.

n.a. = not applicable.

Appendix Table 2.
Matched 1040s, by Filing Status and Household Composition of Constructed Tax Unit in Tax Year 2006

(Number, in millions)

	CPS Characteristics			
Filing Status on 1040	Unmarried With No Dependents	Married	Unmarried With Dependents	Total
Single	40.0	3.9	4.1	48.0
Married Filing Jointly	2.8	52.3	1.0	56.1
Head of Household	5.2	4.1	10.1	19.4
Other	0.6	1.8	0.3	2.7
Total	48.6	62.1	15.5	126.2

Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns.

The sample is restricted to 1040s filed by taxpayers who are not claimed as dependents by other taxpayers.

Appendix Table 3.
Distribution of Individuals and Exemptions, by Filing Status on 1040 and Household Composition of Constructed Tax Unit in Tax Year 2006

(Number, in millions)

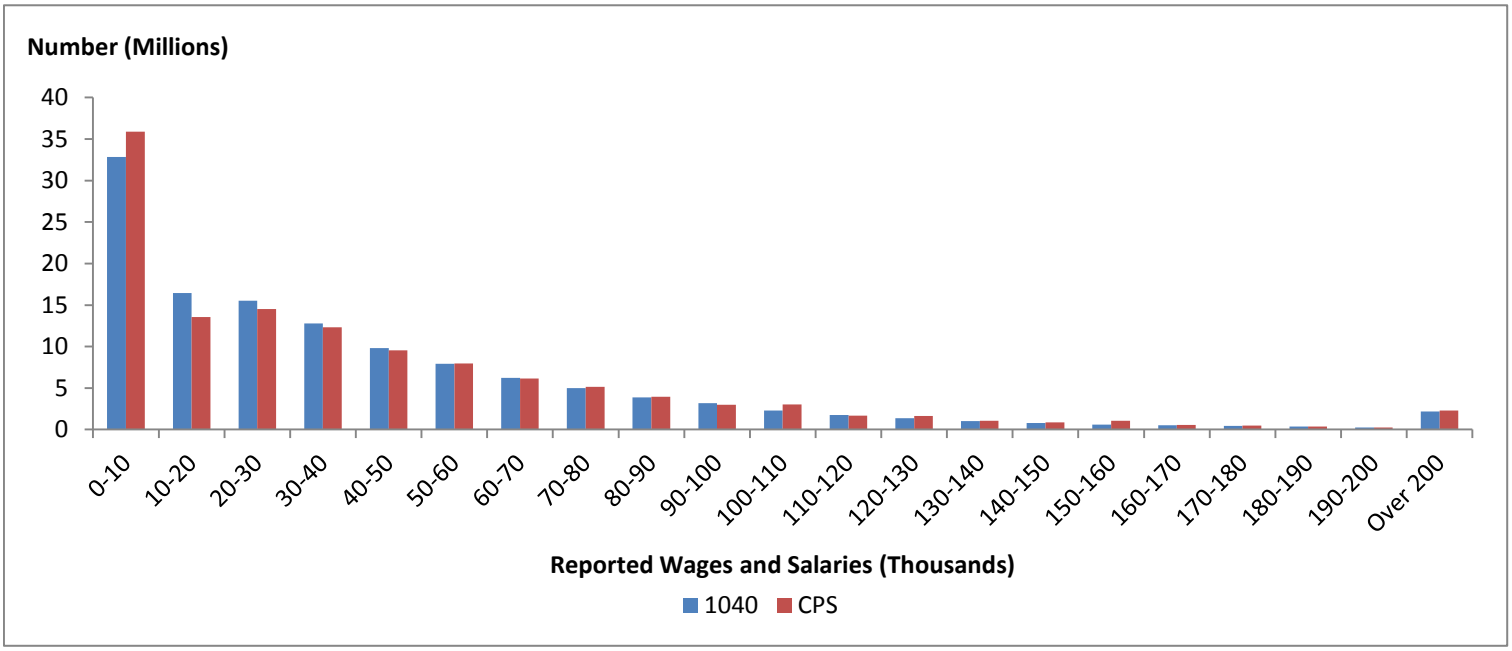
Number of 1040s Matched to Constructed Tax Unit and Filing Status	CPS Characteristics: Unmarried With No Dependents		Married		Unmarried With Dependents		All	
	Individuals	Exemptions	Individuals	Exemptions	Individuals	Exemptions	Individuals	Exemptions
Only One 1040 Matches								
All members linked to the same 1040								
Single	38.5	38.2	0.4	0.5	1.2	1.3	40.1	40.0
Married	1.7	4.5	132.9	140.9	1.4	2.0	135.9	147.4
Head of household	4.8	10.0	0.9	1.2	15.0	16.0	20.8	27.2
Other	0.2	0.3	0.4	0.3	0.2	0.2	0.8	0.8
Subtotal	45.3	52.9	134.6	142.9	17.8	19.4	197.6	215.3
Some members do not appear on 1040	n.a.	n.a.	20.9	16.1	11.6	7.5	32.5	23.6
More Than One 1040 Matches								
All members linked to 1040s	1.7	5.8	12.5	15.1	4.4	6.5	18.5	27.4
Some members do not appear on 1040s	n.a.	n.a.	2.3	2.1	1.0	1.0	3.3	3.1
Total	46.9	58.7	170.3	176.2	34.7	34.5	251.9	269.5

Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns.

The sample is restricted to constructed tax units with at least one linked 1040 filed by a taxpayer who is not claimed as a dependent by another taxpayer. The number of exemptions (personal and dependent) claimed is based on information from the 1040. The number of individuals is based on membership in the constructed tax unit.

n.a. = not applicable.

Appendix Figure 1.
Distribution of Tax Units by Wages and Salaries in Tax Year 2006, by Data Source



Source: Author's calculations, using data from the Census Bureau's linked Current Population Survey (CPS) and tax returns.

The sample is restricted to constructed tax units with at least one linked 1040 filed by a taxpayer who is not claimed as a dependent by another taxpayer.